Analysis of World War One Diaries using Natural Language Processing

Ashley Grace Dennis-Henderson

December 2020

Thesis submitted for the degree of
Master of Philosophy
in
Applied Mathematics
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

Contents

List of Tables

List of Figures

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the Universitys digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: ... . Date: 8/12/2020

Acknowledgements

Firstly, I would like to thank my supervisors, Professor Matthew Roughan, Dr Lewis Mitchell and Dr Jono Tuke for their guidance and support in researching and writing this thesis.

I would like to acknowledge the State Library of New South Wales for providing the data which made this research possible. I would also like to acknowledge the ARC Centre of Excellence in Mathematical and Statistical Frontiers who provided support in attending conferences related to my research.

I would like to thank my mum, who has always believed in me and inspired me to do my best. This thesis would not have been possible without her support.

Finally, I would like to thank my family and friends, especially my grandparents, Kristy, Sophie, Chelsea, Jill, and Peter, for their support and encouragement over the past two years.

Disclaimer

This thesis contains words that whilst considered part of the normal vernacular during the time that the diaries were written, today they would be deemed as offensive and insensitive. The inclusion of these words is simply for accuracy in the data and are not included with the aim of causing discomfit or offence to anyone.

Abstract

World War I was a significant event in Australian history and as such it has been extensively researched. The analysis of relevant primary sources has included the close reading of war diaries. Close reading involves reading the diaries to understand what the soldiers went through. With the advancement of computational techniques, we now have the ability to analyse large volumes of text, and this concept is known as *distant reading*. This project focuses on 557 Australian World War I diaries collected and transcribed by the State Library of New South Wales, and aims to use distant reading methods to determine what the soldiers wrote about and how they felt over the course of the war.

In order to perform our analysis over time, we first needed to extract dates from the diaries. This was done using regular expressions. However, some problems were found in the extracted dates, including missing data such as the month or year, dates which were written incorrectly, and dates that were actually within the text referring to events that happened rather than the date the diaries were written. Hence, an optimisation program was formed to fix these problems and give more accurate information about when the diaries were written.

We then considered several types of analysis to understand what the soldiers wrote about, including word frequencies, tf-idf (term frequency - inverse document frequency), and topic modelling. It was found that whilst all three of these techniques gave results that would be expected when considering World War I diaries, they also showed different aspects of the war. In particular, through considering the tf-idf results for 1916 we see many words regarding places and battles in the Middle East. However, when considering topic modelling for this time period we see more words regarding Europe. Sentiment analysis, more specifically dictionary-based methods, was then used to understand the emotions of the soldiers over time. Using our dic-

tionaries, each month was given an overall sentiment score from -1 (very negative) to +1 (very positive). It was found that the average sentiment of the diaries ranged between 0 and 0.2. We were also able to compare this to our topic modelling results to determine which topics corresponded to peaks and dips in our sentiment.

Chapter 1: Introduction

World War I was a significant event in Australian history which helped shape Australia's national identity. Due to its importance, World War I has been extensively researched, and individual war diaries have been analysed through traditional *close reading* methods. However, the digitisation and transcription of the State Library of New South Wales' collection of World War I diaries provides the opportunity to gain new insights into the war by applying mathematical and computational techniques to the collection as a whole. This methodology is known as *distant reading*. Our aim in this thesis was to use mathematical and computational techniques to analyse a set of Australian World War I diaries in order to understand what the soldiers wrote about and how they felt about these topics over the course of the war. More specifically, we aimed to

1. Develop a technique for extracting dates from the diaries,
2. Perform various types of topic analysis including word frequency analysis, tf-idf, and topic modelling, to determine what topics were important to soldiers at specific times in World War I,
3. Perform sentiment analysis to try to quantify the emotions of the soldiers, and compare this with the topics they wrote about.

The use of distant reading, as opposed to traditional close reading, to analyse large text corpora is one of the concepts which form the basis of the emerging field of digital humanities [3]. Close reading involves each individual reading the text in order to comprehensively understand what happened, whilst distant reading uses mathematical and computational techniques to analyse the text, looking for overall patterns [4]. The use of mathematical and computational techniques on text corpora is also known as *natural language processing*. Distant reading has several advantages in that a researcher can analyse large text corpora, create visualisations

of their analysis, and remove some researcher bias [3]. Further, these techniques can be combined so that interesting patterns found though distant reading can be explored further by close reading of the corresponding text.

Various researchers have analysed war diaries through close reading, including Michael Caulfield [5], Peter Cochrane [6, 7, 8] and Bill Gammage [9]. However, distant reading methods have not been applied to these diaries before. Computational analysis has been used on other diaries such as those of Anne Frank [10] and Martha Ballard [11] as well as in the *Memories of War* project to analyse Italian war bulletins [12, 13].

In the following paragraphs, I will review some of the basic terminology used throughout my thesis.

The data for this project was obtained from the State Library of New South Wales, which had collected it as part of the European War Collecting Project. Before this data could be used, it was necessary to clean it. Cleaning the text involved two main steps: creating metadata and extracting dates. *Metadata* is a set of data which gives identifying information about your data. For instance, our metadata includes information such as the author of the diary and when it was written. Dates were extracted from the diaries using a combination of regular expressions and optimisation. *Regular expressions* are rules which specify a set of strings to match within our text [14, 15]. *Optimisation programs* allow us to determine a set of variables which best fit our known data and constraints. In this case, we optimised to find the date which is most likely given the raw extracted date and known constraints. For example, dates must be in chronological order.

Four main techniques are used throughout this thesis to investigate the content of the diaries. These are the consideration of *word frequencies*, *tf-idf* (term frequency - inverse document frequency), *topic modelling* and *sentiment analysis*. The frequency of a word is the number of times that word is used throughout the text. Considering word frequencies is useful in understanding the text overall, but also allows us to investigate how often a particular word or phrase of interest is mentioned. Tf-idf determines the most distinctive words in a document by comparing the frequency of the word within a document with how many documents the word appears in. Words with high frequency in a document that are only used in a small subset

of the documents will be the most distinctive words for that document and hence have a large tf-idf score. Topic modelling is based on the idea that all documents are made up of a series of topics, where topics are a probability distribution over words [16]. Various methods have been developed in order to determine the topics of a corpus, and the distribution of those topics within each document, with the most commonly used method being Latent Dirichlet Allocation (LDA) [17]. Sentiment analysis aims to understand the attitude or emotion of the author towards the subject of the text. Several methods exist to determine this and we will focus on Dictionary Based Methods (DBMs). In DBMs we compare sentiment lexicons with our text in order to determine the average sentiment of the document. A sentiment lexicon is a dictionary of terms with an associated sentiment score.

In Chapter 2 we provide a background on Australia's involvement in World War I, including why Australia supported Britain, the cost of the war in terms of both casualties and finances, and how the war helped create Australia's identity. We also discuss the advantages of distant reading as opposed to the more traditional approach of close reading.

In Chapter 3 we discuss how our data was collected and the necessary cleaning steps, and provide some exploratory data analysis of the collection. The data cleaning was broken into three stages. The first stage involved converting our raw data into a single text file per document with a metadata table giving identifying information for each document. The second stage involved extracting dates from the diaries, with this process being fully explained in Chapter 4. The third stage involved steps such as removing numbers, punctuation, and stop words, as well as singularising words.

In Chapter 4 we explain the steps behind our date extraction process. Dates were initially extracted using regular expressions. However, it was found that there were some problems with these dates, including missing information, mistaken information and non-entry dates dates. To overcome these problems, an optimisation program was created. This program involves parameters which determine the balance of keeping dates close to the raw extracted dates and close to the previous date in the diary. To choose these parameters, we simulate data with each of the problems and run the optimisation program for a variety of parameters to determine the most accurate ones. We then apply our date extraction process to the World War I diaries

using these parameters and discuss why this process is not completely accurate and what we could investigate in the future to improve accuracy.

In Chapter 5 we investigate three methods to understand what the soldiers wrote about in their diaries. To begin, we considered the most frequent words over the entire data set. We categorised the top 100 most frequent words in the diaries as either common, homonyms, or war-related, and found that common words were the most frequent. We also considered the frequency of n-grams related to two chosen subjects of politics and health. We then looked into tf-idf to determine the most distinctive words for each year. Our final method was topic modelling. We discuss the theory behind various types of topic models, including LSA (Latent Semantic Analysis), pLSA (probabilistic Latent Semantic Analysis), LDA (Latent Dirichlet Allocation), and SBMs (Stochastic Block Models). We then used LDA to investigate the topics written about in the diaries over time.

In Chapter 6 we consider sentiment analysis. We discuss the theory behind the three categories of sentiment analysis: dictionary based methods (DBMs), supervised learning methods and unsupervised learning methods. We then use DBMs to analyse our data, using a variety of sentiment dictionaries including: AFINN, ANEW, Huliu, Syuzhet, Mcdonald, NRC, SenticNet, and SentiWordNet.

The code used for this project, the metadata and the simulations results from Section 4.3 can be found in the GitHub repository:

`https://github.com/AshleyDennisHenderson/Analysing-Australian-WW1-Diaries/`

An interactive web app was created to allow users to explore the diaries using the techniques described throughout this thesis. This web app can be located using the following link:

`https://ashleydennis-henderson.shinyapps.io/Analysing-WW1-Diaries/`

Chapter 2: Background on World War I

Word War I, also known as the Great War or the First World War, had a significant impact worldwide. The war began between Austria-Hungary and Serbia, with the assassination of Austrian Archduke Franz Ferdinand on 28 June 1914 [1]. However, due to a network of alliances, on 4 August 1914, Great Britain declared war against Germany after the German invasion of Belgium. The war continued for over four years before the Armistice was signed on 11 November 1918.

The outbreak of war was greeted enthusiastically in Australia as it was in many other nations [18]. Even though Australia had gone through federation in 1901, in 1914 it was still a British Colony. Britain still controlled Australia's contact with countries outside the Empire. Hence, when Britain declared war, Australia was also automatically at war. However, Australia still had a choice as to how much support it would give Britain.

There are several reasons behind Australia's enthusiastic response to war [18]. Firstly, Australia was loyal to the British Empire, with a majority of its immigrants, cultural and political traditions coming from Britain. Also, Britain, and more specifically the Royal Navy, was considered Australia's best defence against growing powers in the region. At the time, Japan's power was growing and Australia's colony of Papua shared a border with the German Empire's outpost of New Guinea. Hence, it was believed that the key to Australia's defence was mutual help. Helping Britain to defeat Germany during the war would prevent Australia from being vulnerable to the German New Guinea. At the time, war was also seen by Australia and many European countries as a "biological struggle for survival among the races of the

world" [18]. War was a chance for Australians to prove themselves.

The reasons that Australia supported Britain are not necessarily the same reasons individuals chose to enlist [18]. It is not known why each person decided to enlist. However, there are many possible reasons including loyalty to Britain, a wish to return home to Britain, a wish to travel, the belief that it would be exciting and short (i.e., that it would be over by Christmas), the wish to avenge fallen family and friends, community pressure, or unemployment. During 1914, Australia was going through drought, and the outbreak of war led to unemployment in some industries, which suggests that some men may have enlisted in order to have regular work with good pay.

Initially, Australia offered Britain a force of 20,000 troops and control over Australia's Navy [18, 19]. This initial contingent of men was filled within weeks, which prompted Australia to offer a further 50,000 men [18]. Australia's troops were involved in conflicts both in the Middle East and on the Western Front, but the most well-known engagement was the Gallipoli campaign. A timeline of Australia's involvement in World War I is given in Appendix A.

World War I took a toll on all countries in many ways, especially in terms of loss of life, and economic impact. There are no accurate numbers in regard to the casualties. However, it is estimated that there were approximately 37,500,000 military casualties, including 8,500,000 military deaths [1]. There is even less certainty regarding the number of civilians who died during the war. However, it has been estimated that there were around 13,000,000 civilian deaths. Table B.1 in Appendix B gives estimates of military casualties based on the U.S. War Department and the Statistical Services Center, Office of the Secretary of Defence of the U.S.A .

During the war, Australia had a population of approximately 4.9 million, of whom 416,809 enlisted with approximately 335,000 of those deployed overseas [20]. This enlistment number represented 38.7% of the male population aged between 18 and 44. Of those who embarked, around 60,000 died, 155,000 were wounded in action, 4,000 were taken as prisoners of war, and there were approximately 431,000 cases of sickness or non-battle related injuries. Overall, Australia had a casualty rate of 64.8% of those who embarked.

With the belief that the war would be short and cheap, Australia also promised to cover the costs of its troops [19]. However, by 1920 the war had cost Australia approximately £376,993,053. Approximately 70% of this was raised through loans and taxation, whilst the rest was owed to the British government for goods and services provided to the Australian army. This cost increased due to interest, and reparation and pension costs to approximately £831,280,947 by 1934.

The war had a massive role to play in forming Australia's identity. Throughout the war, though primarily as a result of the Gallipoli campaign, the ANZAC (Australian and New Zealand Army Corps) legend was formed. This legend is based on the stereotype of the Australian soldier, also known as a 'digger', as "a superb fighter, something of a larrikin, instinctively egalitarian, distrustful of authority, endlessly resourceful, dryly humorous and above all, loyal to his mates" [18]. This list is still seen by many as the set of characteristics of an Australian. The ANZAC legend and the sacrifice of the diggers is still remembered, especially on ANZAC day where Australians

> "remember the commitment and sacrifice of all our men and women who have served in conflicts and on operational service, and those who continue to serve today ... we pause to recommit ourselves to one another, our nation and the ideals of mankind", Dr Brendan Nelson [21].

An important question many historians have researched is: "what did the soldiers go through during the war?" In order to answer this question, researchers have studied diaries, memoirs, and other first-hand accounts from those who served. Primarily, this has been done through close reading of a small subset of accounts. Examples of historians who have relied on the close reading of memoir literature include Michael Caulfield [5], Peter Cochrane [6, 7, 8] and Bill Gammage [9].

With the advancement of computational techniques to analyse large volumes of text, we now have the ability to gain new information from digitised World War I texts. This concept is known as distant reading, as opposed to traditional close reading, and is one of the concepts which form the basis of the emerging field of digital humanities [3, 4]. Close reading involves reading individual texts to gain a comprehensive understanding of the people, events, and themes discussed in the text. By contrast, distant reading uses computational techniques to analyse texts, and looks for overall patterns within hundreds or even thousands of texts. Distant reading has

advantages in that large amounts of text can be considered, computational programs can be used to create visualisations and analyses of the data, and the selection bias from the researcher is reduced. These two concepts of close and distant reading can also be combined. Interesting patterns found through distant reading, can be further examined through close reading of the corresponding text to gain insight into why that pattern occurs.

There are still some limitations to distant reading, as many documents from that era are handwritten and as yet there is no OCR (optical character recognition) software that works reliably on handwritten text. This means that in order to use computational techniques to discover what soldiers went through in World War I, any data must be manually transcribed.

In the next section, we will look at World War I diaries and other documents held by the State Library of New South Wales. We will discuss the necessary steps required to clean this data such that it can be analysed using computational techniques in subsequent chapters.

Chapter 3: Data Collection and Cleaning

This project focuses on war diaries from Australians who served in World War I. Diaries are considered rather than other documents, such as letters and memoirs, as they contain temporal information regarding the sequence in which the text was written. This gives us the ability to analyse content and emotion over time. The data used for this project is the 557 diaries from the State Library of NSW's World War I collection. This chapter describes the data collection, the cleaning we performed on the data set, and some exploratory data analysis.

At the end of World War I, Principal Librarian William Ifould and the Trustees of the library began the European War Collecting Project [22]. The aim of this project was to collect documents from all areas where Australians served, which detailed the experiences and personal feelings of the men in battle. To do this, advertisements were placed in Australian, New Zealand, and British newspapers. Since then, the library has continued to receive documents and this collection now contains 966 diaries, letters, war narratives, photographs, and other documents from approximately 450 individuals. The library has digitised this collection by scanning each document, and then transcribed the largely handwritten documents using crowd sourcing. This gives researchers access to digital (text) versions of the documents.

This collection has previously been studied by Michael Caulfield [5] and Peter Cochrane [6, 7, 8]. Similar collections have also been studied, such as Bill Gammage's book on diaries held by the Australian War Memorial [9]. However, the analysis in these books and articles is based on close reading of a few diaries, rather than quantitative analysis of a large number of diaries. This kind of analysis has been

performed on other diaries such as those of Anne Frank [10] and Martha Ballard [11].

As for using computational techniques on war time text, a group of Italian researchers as part of the *Memories of War* project began using natural language processing techniques to analyse Italian war bulletins [12], and created an interface where researchers can track n-grams over time [13].

3.1 Raw Data

Individual diaries can be found and downloaded from the State Library of NSW's website [23]. However, we obtained the complete collection directly from the library. The raw data is structured such that for each document there is a folder containing JSON (JavaScript Object Notation) files, where each JSON file is a page of the document. The folders have names related to the document such as:

document-alexander-diary-1917-1918-roy-alexander-node-80404

and each page file has a name related to which page it is, for example, *page-2-node-103902*. We note that each folder name contains the author's name (if known), what type of document it is (e.g. diary), an item number (if the author had multiple documents) and when it was written. Further, this title may also include information regarding the receiver (if the document was a letter or telegram), or information about the contents of the document. Separate to this folder structure is another JSON file per document containing the URLs for the transcriptions of each page of the document. The structure of the raw data is shown in Figure 3.1.

JSON is a language independent text format built on a collection of name/value pairs [24]. Within the page files, the text is contained within two name/values pairs: { "body" : { "value": text } }. The JSON files also contain other name/-value pairs indicating the title of the document (including author, type of file, and page number), and URLs to the photograph of the page and transcription tool. An example of one of these JSON files is given in Appendix C.

In order to analyse these diaries over time, it is necessary to place them in a date/entry format such as that shown in Table 3.1. Two stages of cleaning steps are required to convert the raw data into this form, explained in Sections 3.2 and 3.3, and

a third cleaning stage is required to prepare our data from this form for analysis, as explained in Section 3.4.

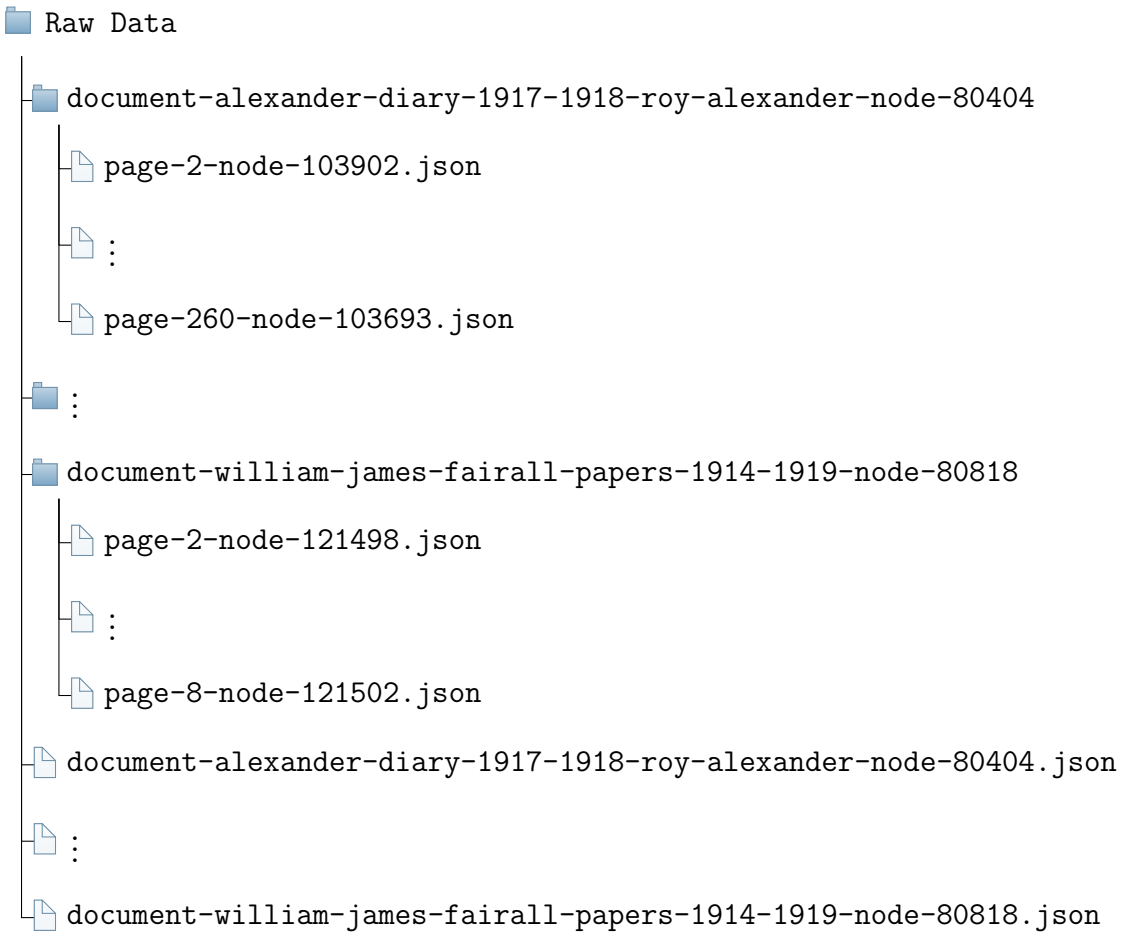


Figure 3.1: Structure of the raw data received from the State Library of NSW.

Date	Entry
02/02/1915	From 8am horses began to arrive alongside, ...
08/02/1915	Was so ill during this period that I was unable ...
01/04/1915	Many large French and British transports ...
05/04/1915	At 5.30a we were towed out from the wharf ...
13/07/1915	At 4pm ship was shifted from Victoria Dock ...
14/07/1915	Did not awaken until 7am, so was too late to ...

Table 3.1: Date/entry format for a selection of entries from Norman Thomas Gilroy's war diaries.

3.2 Data Cleaning: Stage 1

The first stage of data cleaning focuses on converting the raw data into a single text file per document, with a metadata table giving identifying information for that document. The steps discussed throughout this stage are summarised in Figure 3.2.

First, for each document, all page files were opened and the content extracted and combined into a single text file with the same name as the original folder. Note, there were three documents whose page names did not contain numbers, i.e., they were of the form: *page-null-node-130457*, and these page files had to be manually renamed before they were condensed into a single file. These documents were,

- *document-scheidel-letters-received-during-world-war-12-october-1914-10-november-1919-lillian-scheidel-node-80873*,
- *document-series-05-part-01-keith-aubrey-ferguson-correspondence-1915-1918-1933-44-1965-node-80679*,
- *document-vick-letters-1916-1918-frederick-harold-vick-node-80877*.

Using regular expressions various aspects of the document titles were extracted in order to create a metadata table. A regular expression is a rule which specifies a set of strings we wish to match within our text [14, 15]. The metadata variables are described in Table 3.2, and a copy of the metadata table can be found at:

`https://github.com/AshleyDennisHenderson/Analysing-Australian-WW1-Diaries/tree/master/Metadata`

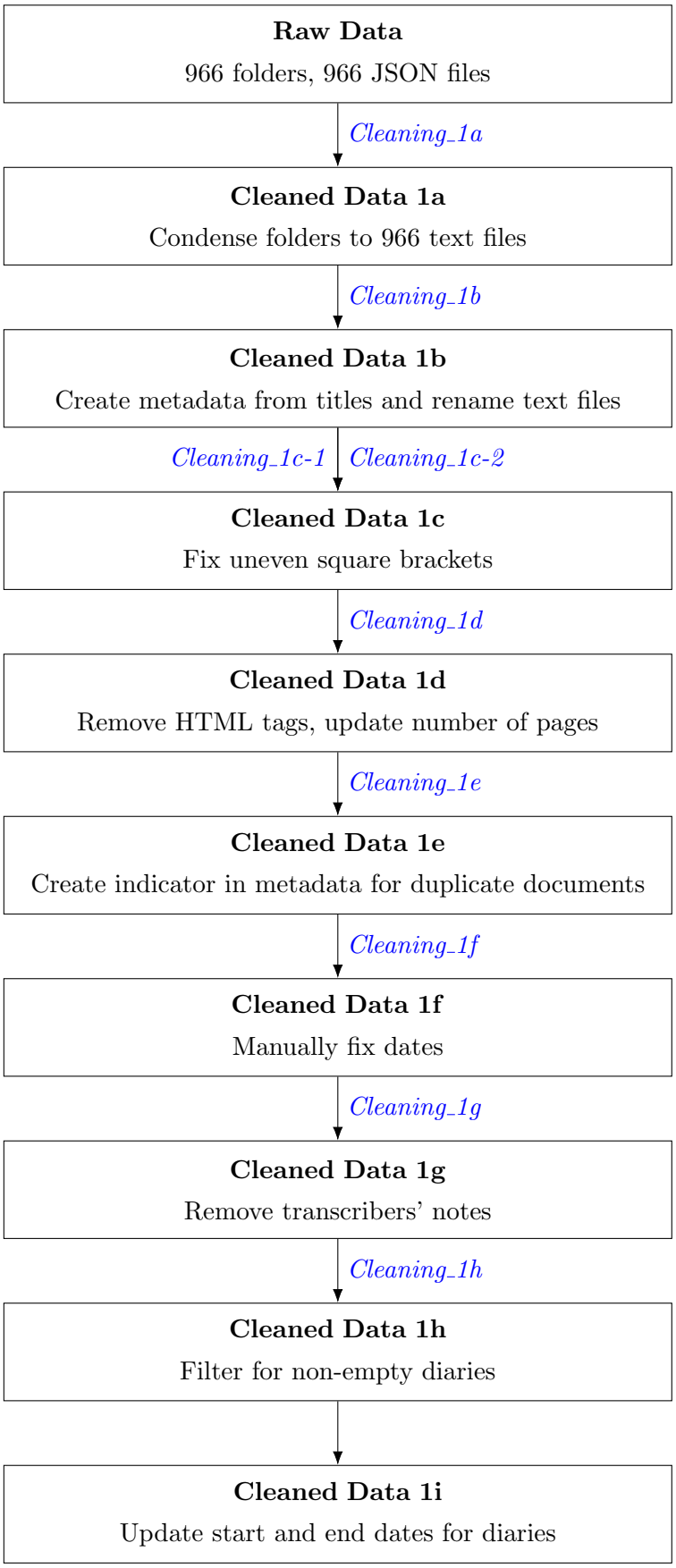


Figure 3.2: Flowchart describing the various cleaning steps in Stage 1. Bold headings are the name of the folder the data was saved in, and italic blue text is the name of the code file used to perform this step.

It was manually confirmed for each document that the extracted information matched the document title, and the metadata table was updated as needed. It is important to note that this metadata table may not be completely accurate as it is solely based on the document title. For instance, if the document title did not contain certain information, such as an author, then this data is not included. It is possible that this information could be found either on the State Library's website or by closer examination of the diaries. However, this was not looked into as every title contained, at a minimum, the item type and an approximate date which is all that is necessary. Also, a series of letters may be written by two authors, but only one is noted, and dates, for items such as memoirs, would most likely refer to when the memoir was written, not the time it was being written about.

All documents were then renamed using the simple format *Document_*.txt*, where * is a different number for each document. Note that these numbers do not give us any information about the document, and are instead based on the file's position when considering the original document names in alphabetical order.

All variables included in Table 3.2 are extracted from the original document titles except Document Name, Number of Pages, Number of Words and Duplicate. The number of pages is based on counting the number of strings of the form *[Page *]* within the text, while the number of words in the document and whether it is a duplicate is determined in future cleaning steps.

These documents contain transcriber's notes which have to be removed before performing our analysis. A transcriber's note is a comment left by the transcriber and is not part of the original document. All transcriber's notes should ideally be surrounded by square brackets. However, as no checks were performed on the number of square brackets during the transcription process, we find that some documents have an uneven number of square brackets. For example, a transcriber may have started their note with a square bracket but forgotten the closing square bracket. We wish to remove this text, hence the next step in our cleaning process is to check which files have an uneven number of square brackets and manually fix them. In most cases this could be done without comparing to the original diary page, for instance, when a bracket was missing from around text such as "indecipherable", "?" or a page number, or when there were two brackets instead of one, or a left bracket instead of a right (or vice versa). However, for cases where it was difficult

Variable	**Type**	**Description**
Document Name	str	Title of the renamed document in the form *Document_**, where * is a number from 1 to 966
Original Title	str	Title of the document before renaming
Document Type	str	letter, letter diary, diary, narrative or other
Author First Name	str	First name of author
Author Last Name	str	Last name of author
Item Number	int	Some authors have multiple documents, the item number distinguishes them. Documents without an item number are given the item number 1.
Start Month	int	Month the author began writing the document 0 = no data 1, . . . , 12 = January, . . . , December
Start Year	int	Year the author began writing the document 0 = no data
End Month	int	Month the author finished writing the document 0 = no data 1, . . . , 12 = January, . . . , December
End Year	int	Year the author finished writing the document 0 = no data
Receiver	str	Name of receiver if one exists
Regarding	str	What the document was regarding, i.e. a place, a battle, a person
Number of Pages	int	How many pages the document contains
Number of Words	int	Number of words the document contains
Node	int	Node number from original title
Duplicate	int	Is there an identical document in the data set 0 = no identical documents 1 = first of a pair of identical documents 2 = second of a pair of identical documents

Table 3.2: Metadata variables and their descriptions.

to separate the transcriber's note from the diarist's text, photos of the original documents were consulted. It was found that 296 documents (around 30%) of the data set required brackets to be fixed, and a list of these documents can be found in the file `Edited_Brackets.csv`, which is available at

`https://github.com/AshleyDennisHenderson/Analysing-Australian-WW1-Diaries/tree/master/Metadata`

Whilst performing the steps above, inconsistencies were noticed in some of the documents. First, we noted that some page numbers were not fully enclosed in square brackets and some had spaces between the square brackets and the page number. Hence our page count is incorrect. This is updated by recounting the number of strings of the form *[Page *]*, including the possibilities that there is a space before the word Page or after the number. We also noticed that some diaries have all dates within square brackets. This is because the dates were already printed in the diary when the soldier bought it, and hence, as the diarist did not write it, the transcriber has added it as a note rather than the diarist's text. This creates a problem as we wish to remove transcriber's notes. However we wish to keep these dates in those diaries. This will be addressed in a future step. Also, the names of days and months in some diaries are written in French, as these diaries were bought whilst the owner was serving in France. Therefore, when extracting dates in Section 3.3, we extract day and month names written both in English and French. Examples of diaries with printed dates, printed dates in French, and no printed dates are shown in Figures 3.3, 3.4, and 3.5, respectively. Finally, we note that some documents have text crossed out, either by the author or a censor. There are two methods used in the transcriptions to show that this text is crossed out. One method is by using the HTML tags for crossed out text (⟨strike⟩, ⟨/strike⟩) and the other is a transcribers note such as "[crossed out by Censor]". However, these notes are not all the same and do not indicate what is crossed out in a way a computer can easily tell. Hence, we will leave all crossed out text in the documents.

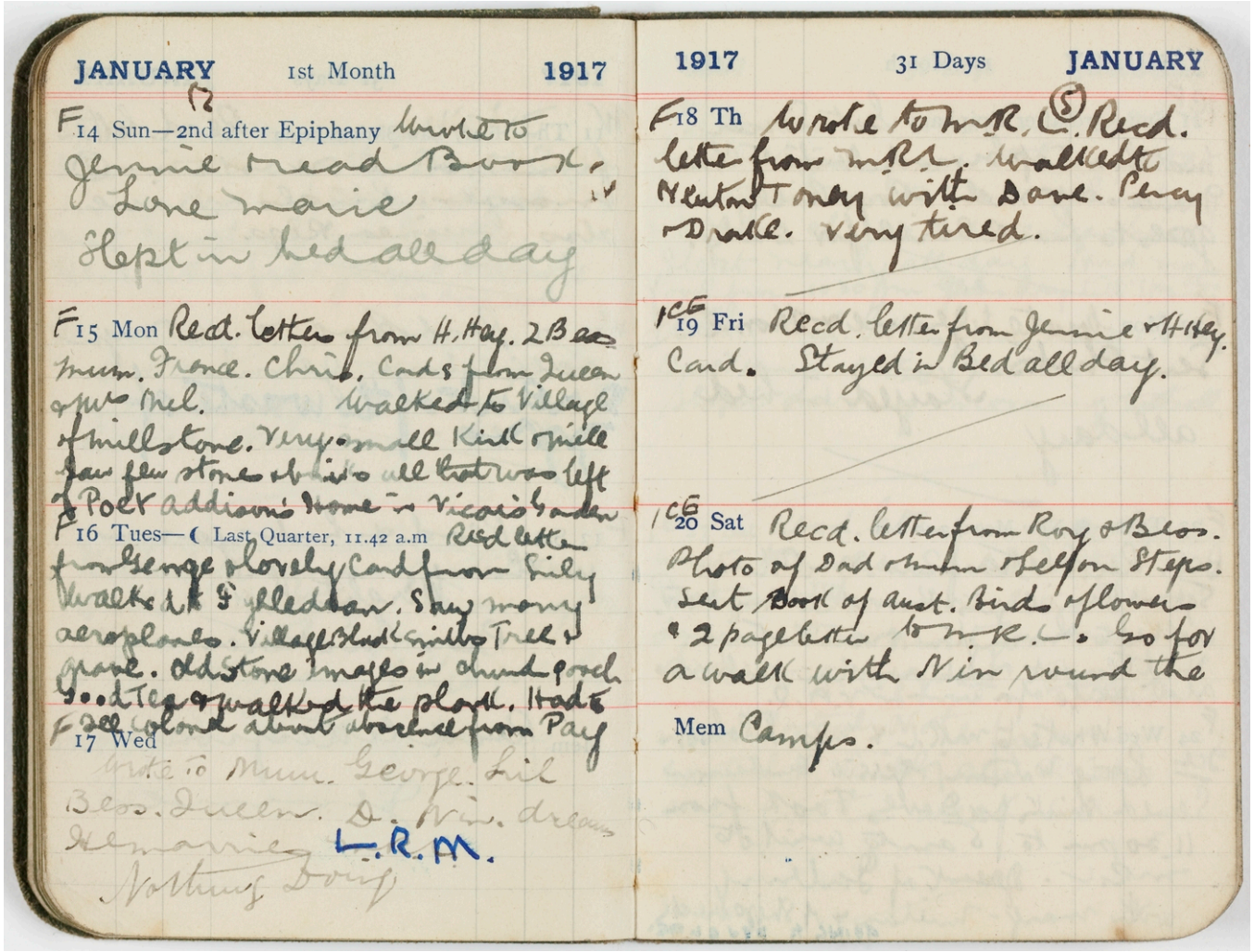
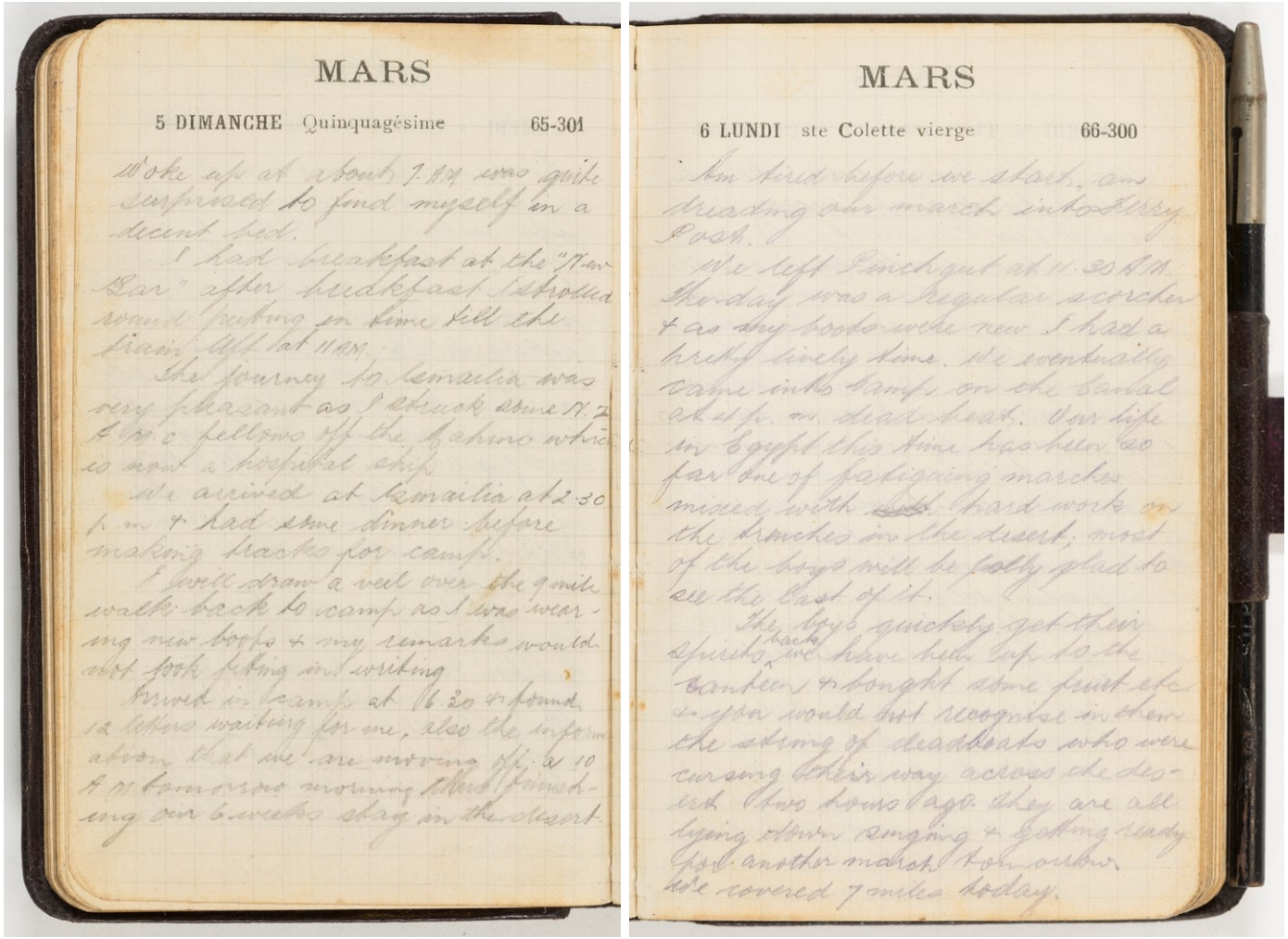


Figure 3.3: Pages from Henry Nicholls' Diary. Note that the dates are printed in this diary.



MARS

5 DIMANCHE Quinquagésime 65-301

Woke up at about 7 AM was quite surprised to find myself in a decent bed.

I had breakfast at the "New Bar" after breakfast I strolled round putting in time till the train left at 11 AM.

The journey to Ismailia was very pleasant as I struck some N.Z. & Aus fellows off the Kahine which is now a hospital ship.

We arrived at Ismailia at 2.30 P.M. & had some dinner before making tracks for camp.

I will draw a veil over the 9 mile walk back to camp as I was wearing new boots & my remarks would not look fitting in writing.

Arrived in camp at 6.20 & found 12 letters waiting for me, also the information that we are moving off at 10 A.M. tomorrow morning thus finishing our 6 weeks stay in the desert.

MARS

6 LUNDI ste Colette vierge 66-300

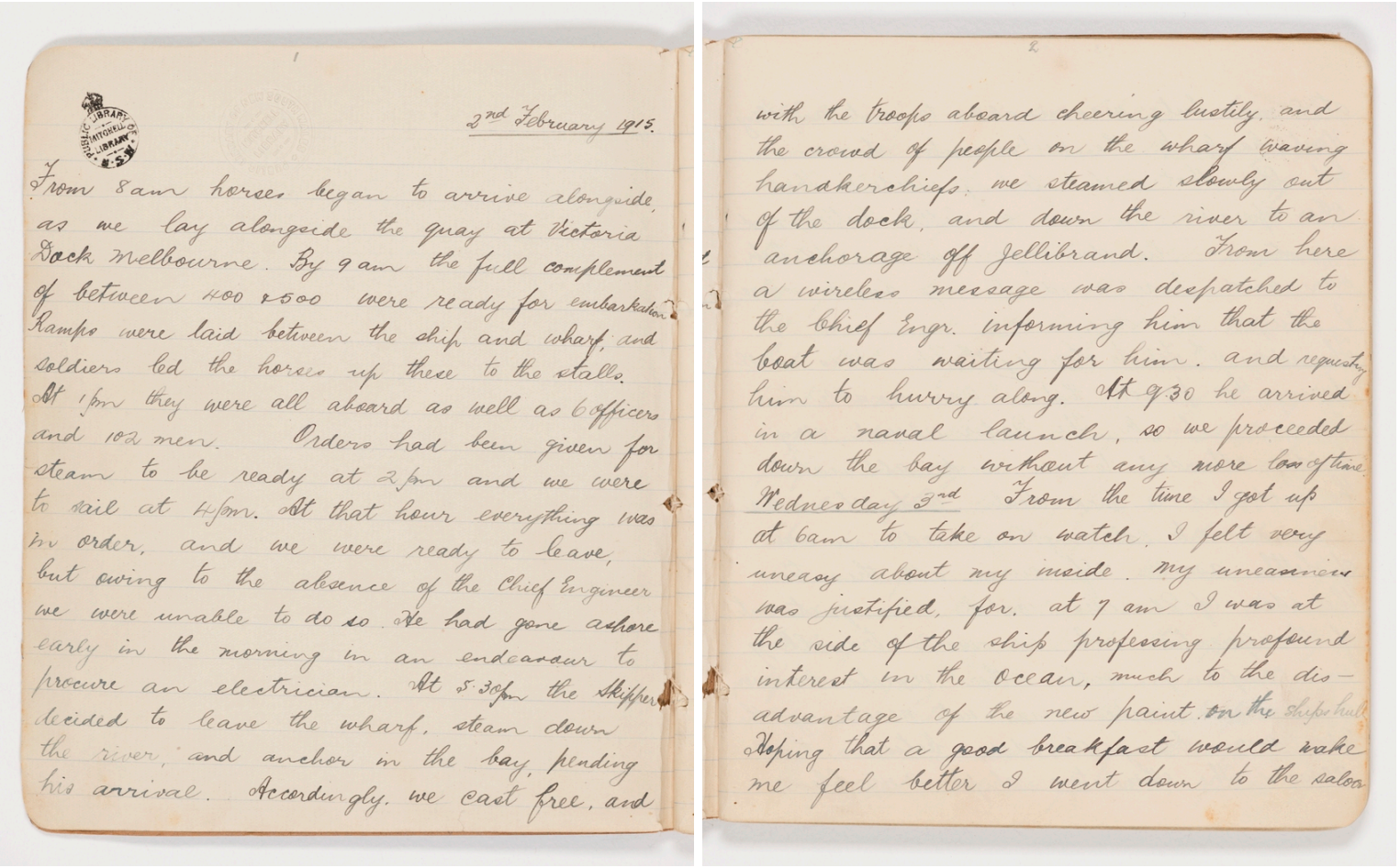
Am tired before we start, am dreading our march into Kirky Post.

We left Serapeum at 11.30 AM. Monday was a regular scorcher & as my boots were new I had a pretty lively time. We eventually came into camp on the Canal at 4 P.M. in dead heat. Our life in Egypt this time has been so far one of fatiguing marches mixed with hard work on the trenches in the desert, most of the boys will be jolly glad to see the last of it.

The boys quickly get their packs, we have been up to the canteen & bought some fruit etc & you would not recognise in them the string of deadbeats who were cursing their way across the desert two hours ago. They are all lying down singing & getting ready for another march tomorrow.

We covered 7 miles today.

Figure 3.4: Pages from Leslie Stuart's Diary. Note that the dates are printed in French in this diary.



2nd February 1915.

From 8 am horses began to arrive alongside as we lay alongside the quay at Victoria Dock Melbourne. By 9 am the full complement of between 400 & 500 were ready for embarkation. Ramps were laid between the ship and wharf, and soldiers led the horses up these to the stalls. At 1 pm they were all aboard as well as 6 officers and 102 men. Orders had been given for steam to be ready at 2 pm and we were to sail at 4 pm. At that hour everything was in order, and we were ready to leave, but owing to the absence of the Chief Engineer we were unable to do so. He had gone ashore early in the morning in an endeavour to procure an electrician. At 5.30 pm the Skipper decided to leave the wharf, steam down the river, and anchor in the bay, pending his arrival. Accordingly, we cast free, and

with the troops aboard cheering lustily, and the crowd of people on the wharf waving handkerchiefs, we steamed slowly out of the dock, and down the river to an anchorage off Jellibrand. From here a wireless message was despatched to the Chief Engr. informing him that the boat was waiting for him, and requesting him to hurry along. At 9.30 he arrived in a naval launch, so we proceeded down the bay without any more loss of time.

Wednesday 3rd From the time I got up at 6 am to take on watch, I felt very uneasy about my inside. My uneasiness was justified, for, at 7 am I was at the side of the ship professing profound interest in the ocean, much to the disadvantage of the new paint on the ship's hull. Hoping that a good breakfast would make me feel better I went down to the saloon

Figure 3.5: Pages from Norman Thomas Gilroy's War Diary. Note that there are no printed dates in this diary.

We then removed any HTML tags from our documents using Python's Beautiful Soup library [25].

Some of the original document titles are very similar to each other, i.e., same author, document type and dates. It is possible that these documents are identical, and hence these documents were checked to ensure that they were not duplicates. The document pairs with similar titles are (1, 92), (66, 140), (351, 836), (369, 846), (488, 586), (491, 835), (681, 711), and (683, 684). However, it was found that none of them were identical. This is because some of the duplicate documents were empty, whilst others only included portions of the other document, meaning that whilst they were similar they were not identical.

As noted earlier, in some documents the dates are within square brackets and hence these brackets must be removed. For each document we extract the text within square brackets and determine the percentage of that text which are month names or days of the week. For documents where this percentage was greater than 1% we then manually checked the document and removed square brackets around dates. This was required for documents 308, 367, 482, 495, 524, and 533. It is still possible that some diaries have problems with the dates which will stop them from being extracted, and hence, when extracting dates in Section 3.3, we check any diaries where there is a low number of extracted dates.

Transcriber's notes within all documents are now removed and the number of words in each document is determined. We then reduce our data set to just non-empty diaries, giving a final data set of 557 diaries. In order to extract dates in Section 3.3 it is necessary to know both the month and year for the start and end dates of the diaries. Hence, for diaries without all of these values the metadata is updated by looking at the State Library of NSW's online catalogue.

3.3 Data Cleaning: Stage 2

Stage 2 of data cleaning involves extracting the dates from the text and creating a date/entry table. An explanation of the method used to extract dates is given in Chapter 4. This method gives us a data table where each row is an extracted date,

and the columns give us the character positions of the first and last character of the raw date in the text, as well as, the day, month and year values of the date. Using this, we then create our date list by taking the date values from this data table, and the entry as the text between the dates. An example of this is shown in Figure 3.6.

Document

E P Bryan [10] 17/8/14 [17] I was swarn in for the 1st A.I.E.F at the Richmond drill hall Vic Aust [89] 18/8/14 [96] I had a very busy day saying good by to all my friends before leaving for Broadmeadows Camp it is a small camp about miles out of Melbourne [237] 19/8/14 [244] We mustered at Sturt St South Melbourne of 1am We had to carry our dinner.

⇓

Extracted Dates

Start Character	End Character	DD	MM	YYYY
11	17	17	08	1914
90	96	18	08	1914
238	244	19	08	1914

⇓

Date/Entry Table

Date	Entry
17/08/1914	I was swarn in for the 1st A.I.E.F at the Richmond drill hall Vic Aust
18/08/1914	I had a very busy day saying good by to all my friends before leaving for Broadmeadows Camp it is a small camp about miles out of Melbourne
19/08/1914	We mustered at Sturt St South Melbourne of 1am We had to carry our dinner.

Figure 3.6: Example of creating a date/entry table from a document using Edward Bryan's war diary. Note that in the document the dates are highlighted in blue, with the superscript numbers indicating the number of characters the start and end of the date is from the beginning of the text.

3.4 Data Cleaning: Stage 3

Stage 3 of data cleaning prepares the entries from our date/entry table to be analysed. This involves changing all text to lowercase, removing numbers and punctuation, singularising words, changing abbreviations and, for some analysis, removing stop words.

Singularising words is the process of removing plurals from words, for example, "kills" becomes "kill" through singularisation. This process is applied to ensure that when considering word frequencies the computer recognises they are the same word. Singularisation is implemented through the `pluralize` package [26] in R.

We have also created a list of commonly used abbreviations, given in Appendix D, which we convert back to the full word before analysis. Once again, this is to ensure that the abbreviation and full word are considered the same when performing our analysis. These abbreviations were found by considering a list of military abbreviations and from observing them in the text.

Stop words are words such as "and", "the", "at", which do not give any information regarding what the text is about. However, whilst stop words such as "not" do not give any information when performing topic analysis, they do give information when doing sentiment analysis. Hence, stop words are removed for topic analysis in Chapter 5,but not for sentiment analysis in Chapter 6. Our stop words list is based on the `stop_words` data set in the `tidytext` package [2] in R. However, we have also manually add stop words to this list. These manually added stop words include day of the week and month names (in both English and French), as well as date suffixes, and the words "html" and "amp". A full list of these added stopwords can be found in Appendix D.

There are two other steps which could be performed to prepare our data for analysis. These are stemming and checking for spelling errors. Stemming is the process of reducing words to their base form. For example, "kill", "kills", 'killing", and "killed" all have the same base word "kill". Both of these steps could be considered in future analysis.

3.5 Summary Statistics

Overall, this data set has 966 documents, with over 95,000 pages and 15,000,000 words. A breakdown of this data set in terms of each type of document is given in Table 3.3. We focus on the 557 non-empty diaries in this thesis.

Type	Number	# Pages	# Words	# Authors
Diary	577	60,004	9,266,353	236
Letter	183	18,497	3,029,163	141
Letter-Diary	22	3,955	639,184	16
War Narrative	32	2,370	624,618	28
Other	152	10,418	2,159,348	111
Total	**966**	**95,244**	**15,718,666**	

Table 3.3: Number of each type of document, along with the number of pages, words and authors. The "Other" category includes documents such as telegrams, photos, postcards, scrapbook, journal articles, and newspaper clippings. Note, there are a total of 577 diaries in this collection, however, only 557 of them are non-empty.

It has been found that the frequency of a word in a text is inversely proportional to a power of its rank in the text, where the most frequent word has rank 1, second most frequent has rank 2 and so on. That is, if f is the frequency of our word, and r is its rank then for some α we have $f \propto r^{-\alpha}$. For most text $\alpha \approx 1$, and this is referred to as Zipf's Law [27]. Mandelbrot generalised this further, giving the Zipf-Mandelbrot law: $f \propto (r + \beta)^{-\alpha}$ [28]. We fit Zipf's law to our data using the `lm` function [] in R, and the Zipf Mandelbrot using least-squares. Figure 3.7 gives us the log-log graph of rank versus frequencies for our words, with lines showing the fitted Zipf and Zipf-Mandelbrot laws. From this graph we note that the relationship between term frequency and rank follows the same path for each type of document. Further, Zipf's law fits well for our higher-ranked words, whilst Zipf-Mendelbrot fits better for our lower-ranked words.

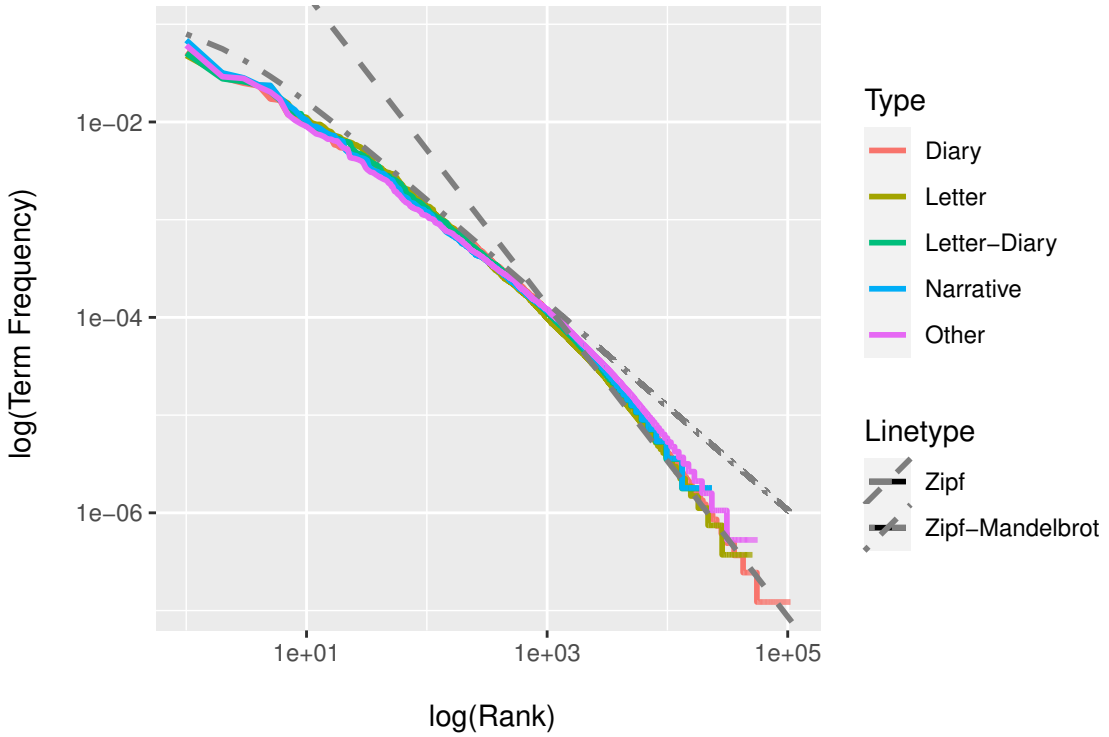


Figure 3.7: Log-log graph of the rank of each word versus their frequency. The Zipf and Zipf-Mandelbrot distributions for our data are also shown.

We now only consider diaries and look at summaries of the number of pages and words the diaries have. From Figure 3.8 we see that both the number of pages and number of words within the diaries is right skewed. From Table 3.4 we note that the median number of pages in a diary is 88, whilst the median number of words is 13,1044. From Figure 3.8 we also note that there appears to be outliers in both graphs, one with around 500 pages and one with around 200,000 words. These outliers belong to the same diary by Sir Charles Rosenthal who wrote a 508 page, 199,141 word diary from September 1914 to December 1918.

	Min	**1st Qu.**	**Median**	**3rd Qu.**	**Max.**
Number of Pages	1	52	88	139	508
Number of Words	109	7,006	13,104	21,437	199,141

Table 3.4: Five number summaries for the number of pages and number of words in the diaries.

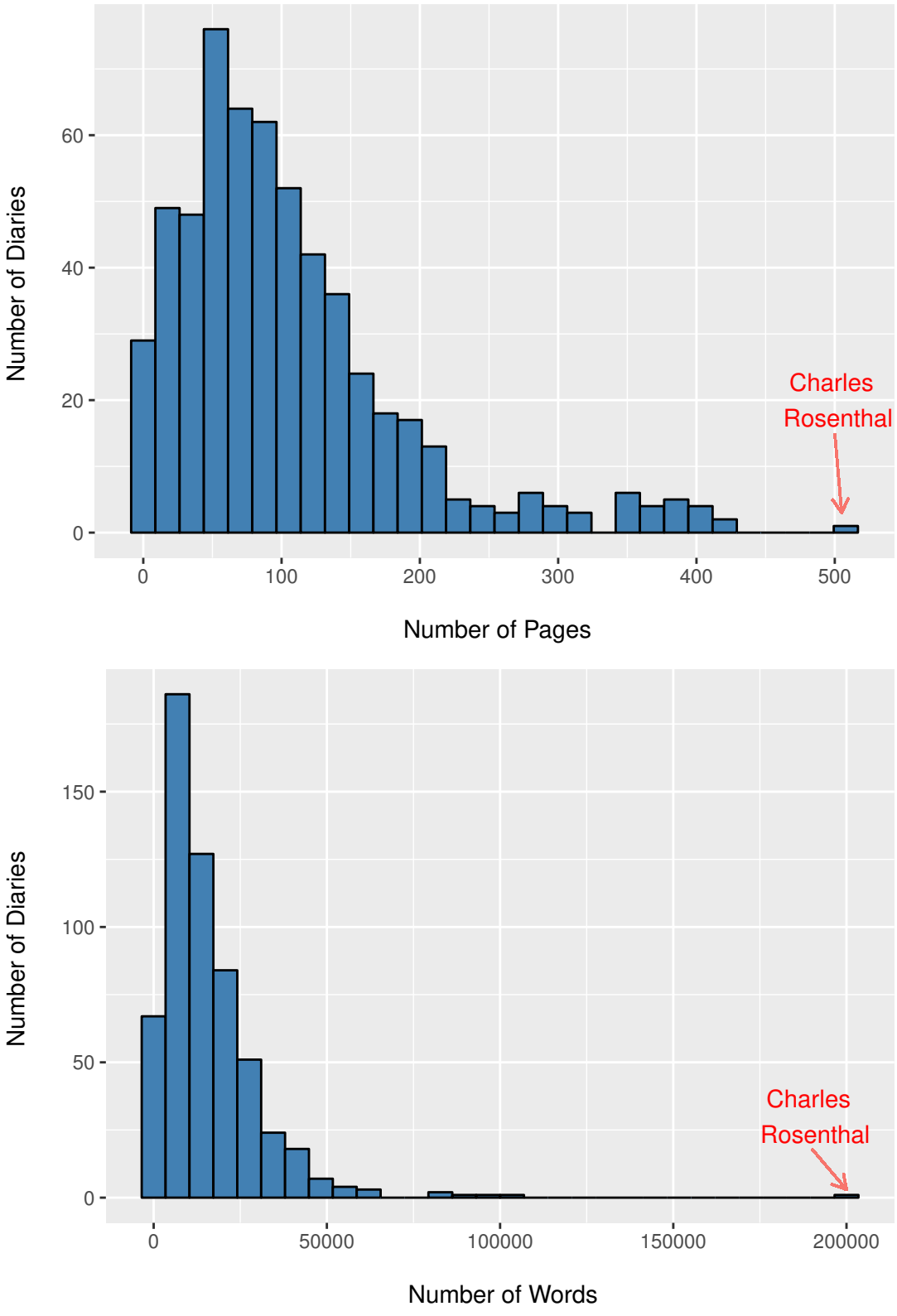


Figure 3.8: Histograms of the number of pages and words in the diaries.

We also consider how many diaries, and the length of diaries an author writes. From Figure 3.9 we see that the number of diaries, pages and words written by diary authors is right skewed. From Table 3.5 we note that the median number of diaries an author writes is two, with majority of the authors writing between one and three diaries. We also note that the median number of pages an author writes is 61, whilst the median number of words is 10,704. From Figure 3.9 there appears to be two outliers in all three histograms. These outliers are Aubrey Wiltshire who wrote 26 diaries, with a total of 2,465 pages and 495,170 words spanning August 1915 to September 1919, and Archie Barwick who wrote 16 diaries, with a total of 2,538 pages and 387,780 words, spanning August 1914 to January 1919.

	Min	**1st Qu.**	**Median**	**3rd Qu.**	**Max.**
Diaries	1	1	2	3	26
Pages	1	61	167.5	347.8	2,538
Words	109	10,704	23,040	52,766	495,170

Table 3.5: Five number summaries for the number of diaries, pages and words written by diary authors.

In the next chapter we will explain our date extraction method.

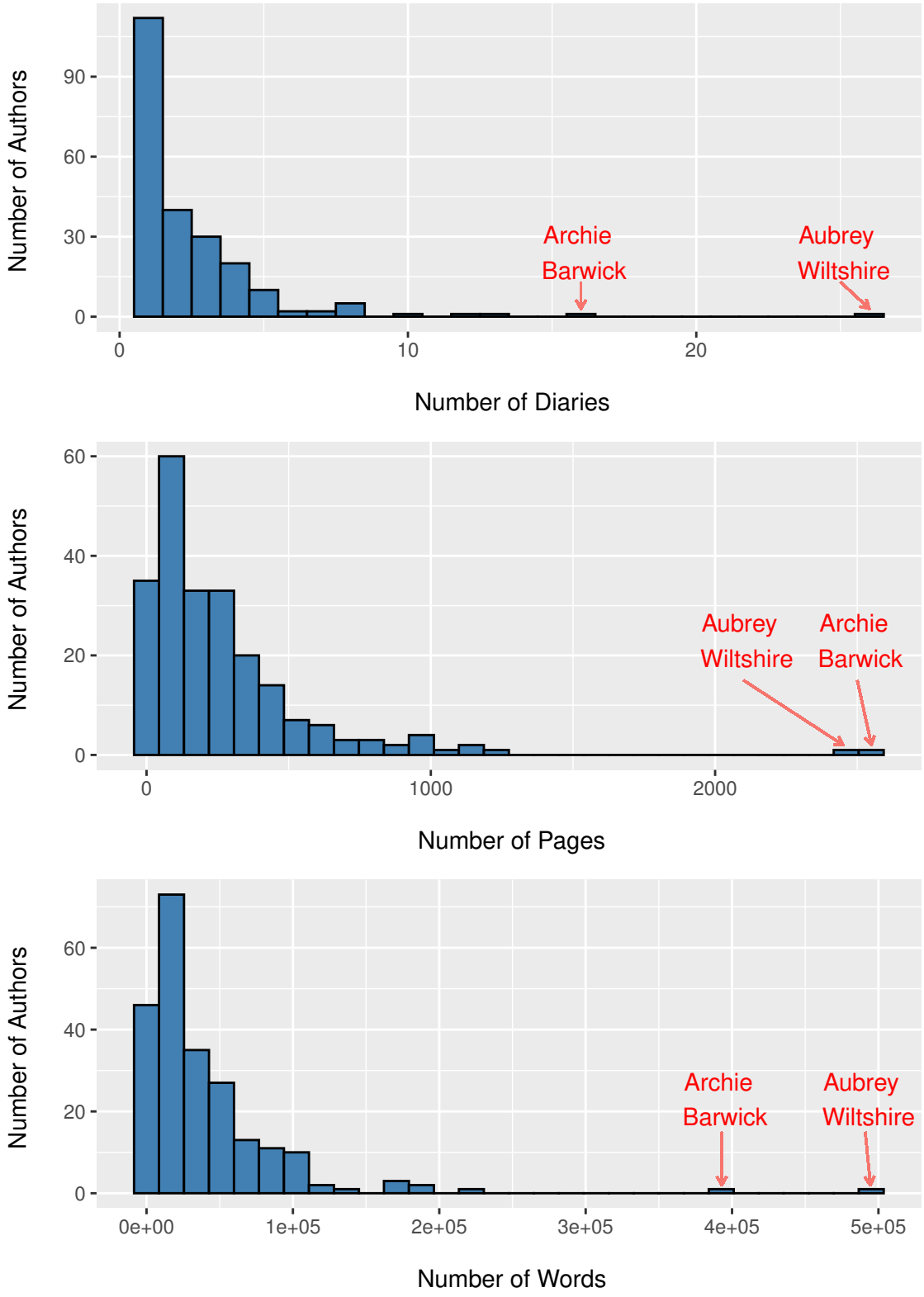


Figure 3.9: Histograms of the number of diaries, pages and words per diary author.

Chapter 4: Date Extraction

An advantage of analysing diaries, rather than documents such as memoirs or war narratives, is that they contain information about the date when the text was written. This means that, rather than analysing the text as a whole, the diaries can be analysed over time. In order to perform our analysis in this way, we require the data to be in a date/entry format such as that given in Table 3.1, meaning that we must first extract the dates and their location within the text. In this chapter we introduce a new method based on regular expressions and optimisation to achieve this.

There are several existing packages to extract dates in Python, such as, `datetime` [29], `date-extractor` [30], `dateutil` [31], `dateparser` [32], and `datefinder` [33]. However, these packages are not ideal for a variety of reasons. Some of them, like `date-extractor`, do not give the location of the date in the text, whilst others, like `dateutil` and `dateparser`, only provide the string surrounding the date as the location. It is necessary to have a known location for our dates in order to find the corresponding entry. Some packages, such as `datefinder` and `dateparser`, will produce too many false positives. For instance, they interpret "5060 tonnes" as a date with the year 5060. Further, `dateutil` will only work on text with a single date, whilst `datetime` requires you to know the format of the dates for which you are looking.

In order to extract the dates (with locations) from our diaries, regular expressions (REs) were used. However, there are difficulties in doing this due to the various ways dates were written within the diaries. Some examples of dates from within the text are given below, with further examples given in Appendix E. Note that the dates are highlighted in blue.

- **Document 15:** "Saturday, April 8th: Company and rifle drill in morning. ... April 9th: Early morning parade, also Church parade. ... "

- **Document 25:** "April 1915 1 Thursday Saw B Sqn AHQ swim their horses Gardens ... 2 Fri Rode back with Clark ... "
- **Document 89:** "17.4.15 Sill steaming along the Gulf of Suez this ... "
- **Document 524:** "Janvier 19 Dimanche Went to Orchestra Symphony Concert at the Alhambra Theatre. ... 20 Lundi Getting on well. It is touch ..."
- **Document 621:** "Friday 25th Jany 18 Left London for Edinburgh 10/15 a.m. arriving ... Saty 2nd. Left Aberdeen Fri evening at 8/15 p.m. and arrived ... 1918 Feby 8th to 27th. Quiet period. ... "

These examples show that extracting dates is difficult due to the different possible formats of dates, and also because within each format it is possible to use different punctuation, different spacing and different abbreviations of month and day of the week (DOW) names. As we have already noted, some month and DOW names are in French. This is due to soldiers buying diaries whilst they were in France which already had the dates printed on the pages.

Based on examples of dates found within the diaries, several different RE patterns were considered. There are four main components to consider: the DOW, day (e.g. 19th), month and year. We also need to consider whether the month is written in number or word form. Throughout this chapter, we will use the following representations of these components:

- dd is a 1 or 2 digit number representing the day. This may or may not include a suffix of either "th", "st", "rd", or "nd".
- mm is a 1 or 2 digit number representing the month,
- yyyy is a 2 or 4 digit number representing the year,
- month is the written out version of the month, including possible abbreviations and French spellings,
- DOW is the day of the week, including possible abbreviations and French spellings.

When extracting dates we are only interested in keeping the day, month and year information, hence, the DOW is only included in the RE if it is not at the beginning or end of the date.

When writing a month in number form we would have a date such as “17.4.15” or “5/12”, hence the first two REs used were:

- dd-mm-yyyy, dd/mm/yyyy, dd.mm.yyyy or dd:mm:yyyy,
- dd-mm or dd/mm.

Note that we have not included dd.mm or dd:mm as possible dates as these are typical formats for times. We also assume that, when writing dates in this format, the diarists use the commonly used Australian form of day, month, year as opposed to other forms such as month, day, year.

When months are written in word form there are many more possibilities, as the diarists could have written the date components in any order. It is also possible to have written dates with just the day, month and year (3 components), or just the day and month (2 components). There are 24 possible ways to order all four components, 6 ways to order three components and 2 ways to order two components, giving a total of 32 possible combinations. However, by making the DOW and year optional in our RE, and taking into account that any RE starting or ending with the DOW is just a RE of three components, we can reduce this down to 12 REs. These REs are shown in Figure 4.1, with the 12 REs given in the top row, and arrows pointing to the date formats that are also extracted using that RE when DOW and year are optional.

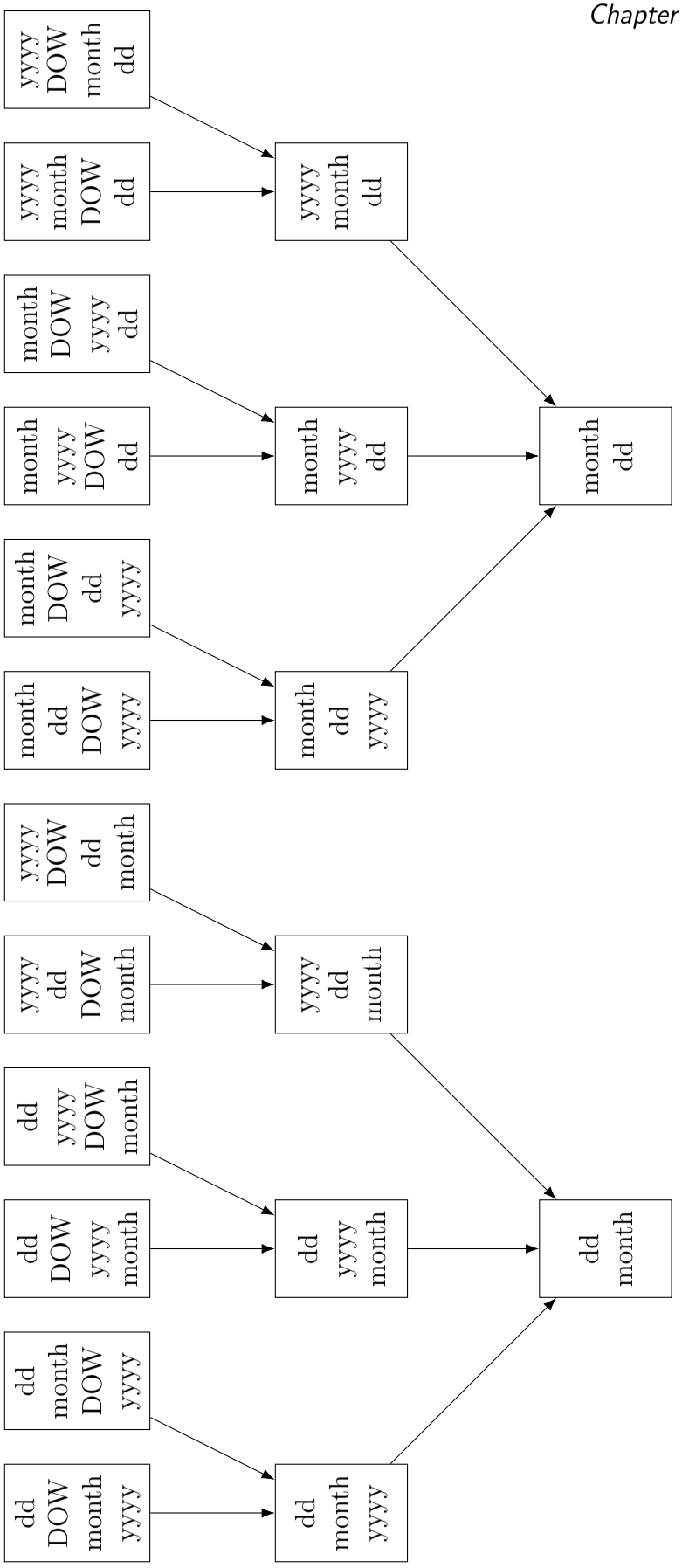


Figure 4.1: Flowchart showing the date formats extracted by our REs when the month is in word form. The top row gives the REs, with arrows pointing to other date formats which would be extracted by this RE when the DOW and year are optional.

Whilst some of these date formats are not likely to occur, we include all of them to ensure as many dates are extracted as possible.

Other possible date formats are the DOW followed by the day, just the day, or just the DOW. However, extracting dates whenever just the day or just the DOW is used can extract many things that are not actually dates. For instance, soldiers may refer to things such as the 1st battalion, or something that has or will happen on a particular DOW, and neither of these are actually dates. Hence, the final we include RE is

- DOW dd

In each of these REs we have also included various punctuation that may have been used throughout the date.

The text extracted using regular expressions will be called our *raw dates*, i.e., the dates are in the exact form they were within the text. Once they are extracted, we convert them into *expanded form* where each date is a vector containing the day, month and year values, i.e., $[dd, mm, yyyy]$. After converting dates to expanded form, three problems were noticed which need to be addressed and these will be discussed in Section 4.1.

4.1 Problems with Extracted Dates

There are three main problems with our extracted dates: missing information, mistaken information, and non-entry dates.

Missing information refers to instances where a diarist has not included the month and/or year when writing the date. This information is usually excluded as it is contained within a previous date and hence, from a human perspective, is is not necessary for it to be included in every date. For instance, if a diarist wrote an entry under the *1st May 1915*, and the next entry is simply under *Tuesday 2nd*, we can assume they are referring to the *2nd May 1915*. Note, it is not possible for a date to have a missing month value without having a corresponding missing year value as dates such as *5th 1915* do not make sense. It was found that approximately

13.91% of the raw dates were missing month values, whilst 53.76% were missing year values. When converting dates from raw to expanded form any missing information is set to zero, and an example of this is given in Table 4.1.

Raw Date	dd	mm	yyyy
1st january 1917	1	1	1917
2nd, jan	2	1	0
3rd. jan	3	1	0
4th. jan	4	1	0

Table 4.1: Example of missing values in our extracted dates from Arthur Hall's diary. Missing values are given the value zero, and are highlighted.

Mistaken information refers to when a diarist has written part of the date incorrectly. There are two cases of mistaken information: when the diarists did not know the correct date, and when the diarist has accidentally written the wrong date. The first case would usually arise when the diarist has forgotten what day it was, for instance, if the soldier was on the front line their priority would not be keeping track of what day it was. In this case, the date is usually one or two days from the correct date. The second case occurs when the diarist knows the date but accidentally writes the wrong date, usually by writing the wrong month down. An examples of writing the wrong month is shown in Table 4.2.

Raw Date	dd	mm	yyyy
apl 1st	1	4	0
apl 2nd	2	4	0
3rd march	3	3	0
4th march	4	3	0
5th april	5	4	0
6th april	6	4	0

Table 4.2: Example of mistaken dates in our extracted dates from Norman Gilroy's diary. Mistaken dates are highlighted.

Extracted dates can be considered to be in one of three categories: entry dates, within-entry dates, and non-dates. Entry dates are those dates written at the start of an entry to indicate what day the entry was written on. Within-entry dates are not the dates the entries were written on, but dates that are referred to within the text. For example, in James Bell's diary he writes:

> "April 9th: Early morning parade, ... Receive another letter from Mother dated April 5th, and write to Jack. ...",

which has a within-entry date of April 5th. Non-dates are those pieces of text which are extracted but are not actually a date, such as *5th battalion*. We are only interested in when the entries were written. Therefore, we wish to remove all within-entry dates and non-dates, which will collectively be called non-entry dates. Examples of non-entry dates are shown in Table 4.3.

Raw Date	dd	mm	yyyy
24th july 1915	24	7	1915
25th july 1915	25	7	1915
5th may	5	5	0
26th july 1915	26	7	1915

Table 4.3: Example of non-entry dates in our extracted dates from Norman Gilroy's diary. non-entry dates are highlighted.

In order for our entries to have the most accurate dates possible these problems need to be fixed, and in order to do so an optimisation program was used.

4.2 Optimisation

An optimisation program was used to fix the problems within extracted dates, as described in Section 4.1.

The aim of the optimisation program is to take the extracted dates (in expanded form) and output dates such that they are as close as possible to the true dates. Note that non-entry dates will be changed to the same date as the previous one, as then they can be removed by simply removing any double up dates. For generalisability, this program was written such that it can be used for any set of dates, rather than specifically for World War I.

Let there be n dates in our set, and let d_i, m_i and y_i be the true day, month and year values, where $i \in \{1, \ldots, n\}$. From our data we also have the following variables:

- $\hat{d}_i$, $\hat{m}_i$ and $\hat{y}_i$, the day, month and year values extracted from our raw dates. These values are zero if not contained in the raw date.
- I_i^d, I_i^m and I_i^y, indicator values for whether a day, month and year were extracted from the raw date. They have value 1 if they were extracted and value 0 if they were not extracted.
- s_y, a start year for the corpus.
- x_i, the number of days since the 1st of January in the start year. This will be known as the *epoch form* of the date.
- s_d, the known start date of the diary in epoch form.

Using these variables, the following optimisation program was formed.

$$\min_{d_i, m_i, y_i} \sum_{i=1}^{n} \Big(\alpha I_i^d |d_i - \hat{d}_i| + 31\beta (I_i^m |m_i - \hat{m}_i|) + 372\gamma (I_i^y |y_i - \hat{y}_i|) \Big) + \sum_{j=1}^{n-1} \delta(x_{j+1} - x_j) + \omega(x_1 - s_d)$$

$$\begin{aligned}
\text{such that} \quad s_d &\leq x_i, \quad \text{for } i = 1, \ldots, n, \\
x_i &\leq x_{i+1}, \quad \text{for } i = 1, \ldots, n-1, \\
x_i &= 372(y_i - s_y) + 31(m_i - 1) + d_i - 1, \\
d_i &\in \{1, \ldots, 31\}, \\
m_i &\in \{1, \ldots, 12\}.
\end{aligned}$$

The objective function includes each of these terms for the following reasons:

- $I_i^d|d_i - \hat{d}_i|$, as we want our optimised day value to be close to the extracted day value if one exists,
- $I_i^m|m_i - \hat{m}_i|$, as we want our optimised month value to be close to the extracted month value if one exists,
- $I_i^y|y_i - \hat{y}_i|$, as we want our optimised year value to be close to the extracted year value if one exists,
- $x_{j+1} - x_j$, as sequential dates should be close together, i.e., it is more likely that two dates are a month apart then a year apart,
- $x_1 - s_d$, as we want our first date to be close to the known start date of the text.

The constraints are included for the following reasons:

- $s_d \leq x_i$, as no date should be before the known start date of the text,
- $x_i \leq x_{i+1}$, as dates should be sequential,
- $x_i = 372(y_i - s_y) + 31(m_i - 1) + d_i - 1$, gives the date as the number of days since the 1st of January in the year s_y, assuming all months have 31 days for simplicity,
- $d_i \in \{1, \ldots, 31\}$, as days can only have values in this range,
- $m_i \in \{1, \ldots, 12\}$, as there are only 12 months.

To demonstrate how the optimisation program should work, we will again consider the examples given in Tables 4.1, 4.2 and 4.3. For missing or mistaken data this program should add/change the data based on the data around it, as demonstrated in Figures 4.2 and 4.3. For non-entry dates this program should change the date to be the same as the previous date as then any double up dates can be removed. This is shown in Figure 4.4.

dd	mm	yyyy
1	1	1917
2	1	0
3	1	0
4	1	0

$\longrightarrow$

dd	mm	yyyy
1	1	1917
2	1	1917
3	1	1917
4	1	1917

Figure 4.2: Example of how the optimisation program works on missing values in our extracted dates from Arthur Hall's diary.

dd	mm	yyyy
1	4	0
2	4	0
3	3	0
4	3	0
5	4	0
6	4	0

$\longrightarrow$

dd	mm	yyyy
1	4	1915
2	4	1915
3	4	1915
4	4	1915
5	4	1915
6	4	1915

Figure 4.3: Example of how the optimisation program works on mistaken dates in our extracted dates from Norman Gilroy's diary. Mistaken dates are highlighted.

dd	mm	yyyy
24	7	0
25	7	0
5	5	0
26	7	0

$\longrightarrow$

dd	mm	yyyy
24	7	1915
25	7	1915
25	7	1915
26	7	1915

Figure 4.4: Example of how the optimisation program works on non-entry dates in our extracted dates from Norman Gilroy's diary.

To implement this optimisation program in Python, Pyomo [34, 35] with Gurobi [36] was used. The program also had to be coded slightly different due to the presence of absolute values within the objective function. Instead it was coded as:

$$\min_{d_i, m_i, y_i} \sum_{i=1}^{n} \left(\alpha I_i^d D_i + 31\beta I_i^m M_i + 372\gamma I_i^y Y_i\right) + \sum_{j=1}^{n-1} \delta X_j + \omega x_1 - \omega s_d$$

$$\begin{aligned}
\text{such that} \quad s_d &\leq x_i, \quad \text{for } i = 1, \ldots, n \\
x_i &\leq x_{i+1}, \quad \text{for } i = 1, \ldots, n-1 \\
D_i &\geq d_i - \hat{d}_i, \\
D_i &\geq -d_i + \hat{d}_i, \\
M_i &\geq m_i - \hat{m}_i, \\
M_i &\geq -m_i + \hat{m}_i, \\
Y_i &\geq y_i - \hat{y}_i, \\
Y_i &\geq -y_i + \hat{y}_i,
\end{aligned}$$

$$
\begin{aligned}
X_i &= x_{i+1} - x_i, \quad \text{for } i = 1, \ldots, n-1, \\
x_i &= 372(y_i - s_y) + 31(m_i - 1) + d_i - 1, \\
d_i &\in \{1, \ldots, 31\}, \\
m_i &\in \{1, \ldots, 12\}.
\end{aligned}
$$

There are five unknown parameters in this optimisation program (α, β, γ, δ and ω). To determine the parameters that give the most accurate dates we use simulations.

4.3 Simulations

To determine the optimisation parameters which give the most accurate dates we created simulated data to represent each possible problem. The optimisation program was then run on each set of simulated data using different parameters, and the optimisation outcomes were compared to the known true dates. A flowchart showing the simulation process is given in Figure 4.5. Each of the steps in this process will be explained in more detail in the following subsections, before considering the simulation results.

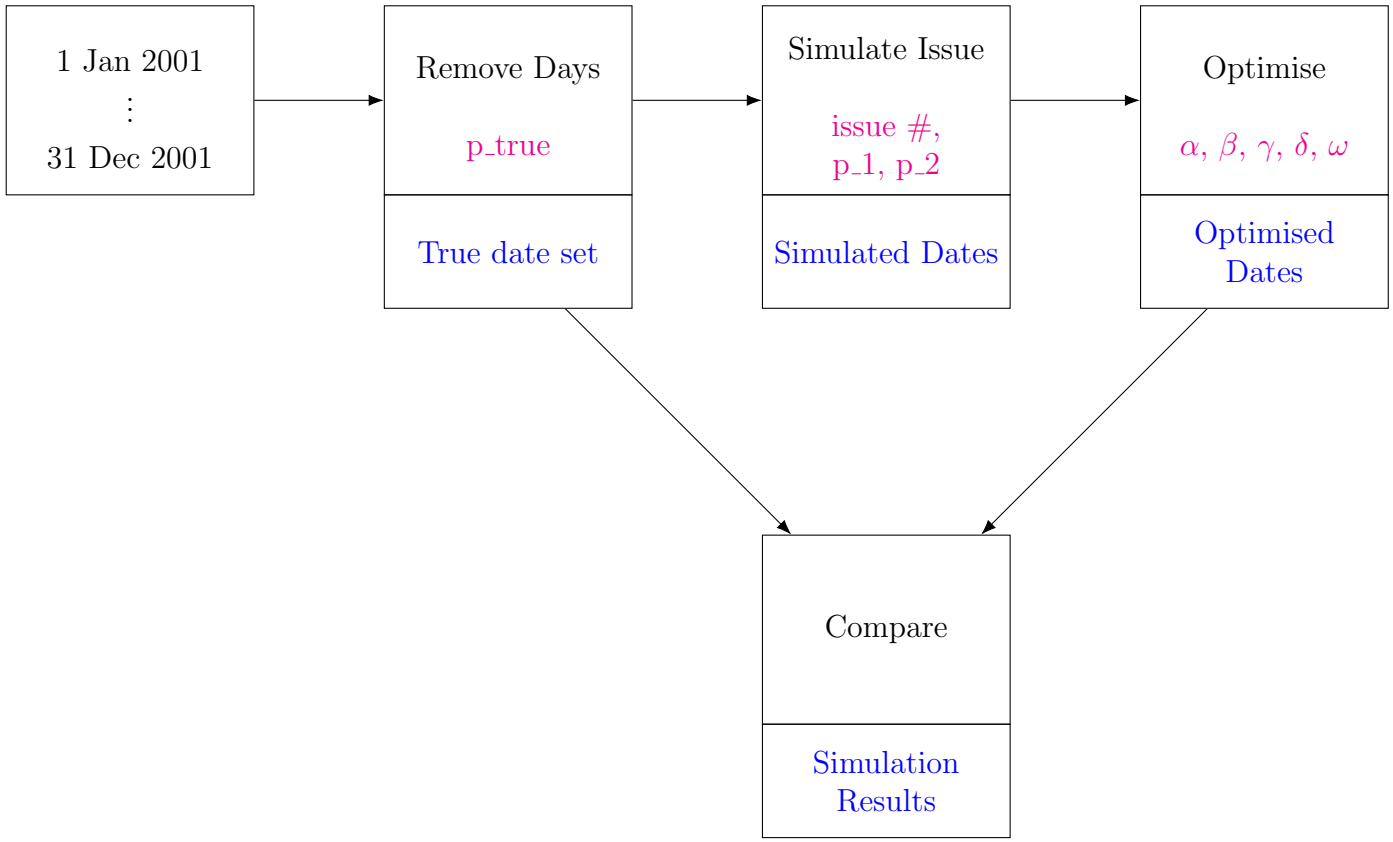


Figure 4.5: Flowchart describing the simulation process. The parameters required for each step are given in magenta, whilst the output of each step is given in blue.

4.3.1 True Date Set

In order to simulate dates and determine the accuracy of the optimisation program, we must have a set of known true dates. To create this set we first choose an entire year of dates and a start year (in order to convert our dates to epoch form). For our simulations we chose all days from 2001, with 2000 as the start year. Note, this choice of dates and start year is unimportant as it will not affect the results. Hence, we have a data frame of dates as given in Table 4.4.

ID	**dd**	**mm**	**yyyy**
1	01	01	2001
2	02	01	2001
3	03	01	2001
⋮		⋮	
373	29	12	2001
374	30	12	2001
375	31	12	2001

Table 4.4: Data frame of all dates from 2001.

It is unlikely that all diarists would write daily entries. Therefore, we need to remove some of these dates. It is also unlikely that a diarist would write on evenly spaced days, i.e., every second or third day. Hence, we simulate a random subset of the dates. Given the proportion of dates we wish to keep, p_{true}, we randomly sample $\lfloor p_{\text{true}} \times 375 \rfloor$ indices (without replacement) from our ID values. The dates corresponding to the sampled ID values become our new true date list, and we update our ID values such that they go from 1 to the length of our list. An example of this process is given in Figure 4.6.

Since it is possible that some diarists wrote entries on more dates than others we wish to consider varying proportions of dates kept. As a result, we consider the values $p_{\text{true}} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Also, as this is a randomised list of dates we consider each proportion 100 times, using seed values in $\{1, \ldots, 100\}$.

ID	dd	mm	yyyy
1	01	01	2001
2	02	01	2001
3	03	01	2001
⋮		⋮	
370	29	12	2001
371	30	12	2001
372	31	12	2001

Random Sample of ID's:
3, 15, 18, . . . , 337, 342, 357

$\longrightarrow$

ID	dd	mm	yyyy
3	03	01	2001
15	15	01	2001
18	18	01	2001
⋮		⋮	
337	27	11	2001
342	01	12	2001
357	16	12	2001

$\downarrow$

ID	dd	mm	yyyy
1	03	01	2001
2	15	01	2001
3	18	01	2001
⋮		⋮	
35	27	11	2001
36	01	12	2001
37	16	12	2001

Figure 4.6: Example of removing random days from the data frame of all dates in 2001 to form our true date set.

4.3.2 Simulation Process

We now create our simulated data sets by altering our true date sets for each possible problem. As it is too complex to consider every possible problem and the combination of problems, we only consider the following eight scenarios:

- Problem 1: Missing Years
- Problem 2: Missing Months and Years (paired)
- Problem 3: Missing Months and Years (not-paired)
- Problem 4: Days randomly off by 1 less
- Problem 5: Days randomly off by 1 more

- Problem 6: Months randomly off by 1 less
- Problem 7: Months randomly off by 1 more
- Problem 8: Inclusion of non-entry dates

The simulation process for each problem is described below.

Problem 1: Missing Years

Given n dates and a proportion of missing values p_1, we simulate missing years by first randomly sampling $\lfloor p_1 \times n \rfloor$ indices (without replacement) from our ID values. We then set the year value to zero for the dates corresponding to the sampled ID values, such as shown in the example in Figure 4.7.

ID	dd	mm	yyyy
1	05	03	2001
2	07	03	2001
3	12	03	2001
4	13	03	2001
5	21	03	2001

$\longrightarrow$

ID	dd	mm	yyyy
1	05	03	0
2	07	03	2001
3	12	03	0
4	13	03	0
5	21	03	2001

Random Sample of ID's: 1, 3, 4

Figure 4.7: Example of simulating missing years.

Problem 2: Missing Months and Years (Paired)

Given n dates and a proportion of missing values p_1, we simulate missing months and years (paired) by first randomly sampling $\lfloor p_1 \times n \rfloor$ indices (without replacement) from our ID values. We then set the month and year values to zero for the dates corresponding to the sampled ID values, such as shown in the example in Figure 4.8.

ID	dd	mm	yyyy
1	05	03	2001
2	07	03	2001
3	12	03	2001
4	13	03	2001
5	21	03	2001

$\longrightarrow$

ID	dd	mm	yyyy
1	05	0	0
2	07	0	0
3	12	03	2001
4	13	0	0
5	21	03	2001

Random Sample of ID's: 1, 2, 4

Figure 4.8: Example of simulating missing months and years (paired).

Problem 3: Missing Months and Years (Not Paired)

Given n dates and two proportions of missing values p_1 and p_2, with $p_1 \leq p_2$, we simulate missing months and years (not paired) by first randomly sampling $\lfloor p_2 \times n \rfloor$ indices (without replacement) from our ID values. This gives us n_2 ID values. We then take the first $p_1/p_2 \times n_2$ of these ID values and set the month and year values of the corresponding dates to zero. For the remaining ID values we set the year value of the corresponding dates to zero. An example of this process is given in Figure 4.9.

ID	dd	mm	yyyy
1	05	03	2001
2	07	03	2001
3	12	03	2001
4	13	03	2001
5	21	03	2001

Random Sample of ID's: 1, 2, 4, 3

↓

ID	dd	mm	yyyy
1	05	0	0
2	07	0	0
3	12	03	2001
4	13	0	0
5	21	03	2001

Random Sample of ID's: 1, 2, 4, 3

⟶

ID	dd	mm	yyyy
1	05	0	0
2	07	0	0
3	12	03	0
4	13	0	0
5	21	03	2001

Figure 4.9: Example of simulating missing months and years (not paired).

Problem 4: Days randomly off by 1 less

Given n dates and a proportion p_1, we simulate days randomly off by 1 less by first randomly sampling $\lfloor p_1 \times n \rfloor$ indices (without replacement) from our ID values. The dates corresponding to the sampled ID values are then altered according to the following. If the date is 1st January, then it is changed to 31st December of the previous year. If the date is the 1st of any other month then it is changed to the 31st of the previous month. For any other date the day value is simply reduced by 1. An example of this process is given in Figure 4.10.

ID	dd	mm	yyyy
1	01	01	2001
2	01	03	2001
3	05	03	2001
4	07	03	2001
5	12	03	2001
6	13	03	2001
7	21	03	2001

$\longrightarrow$

ID	dd	mm	yyyy
1	31	12	2000
2	31	02	2001
3	05	03	2001
4	07	03	2001
5	11	03	2001
6	13	03	2001
7	20	03	2001

Random Sample of ID's: 1, 2, 5, 7

Figure 4.10: Example of simulating days randomly off by 1 less.

Problem 5: Days randomly off by 1 more

Given n dates and a proportion p_1, we simulate days randomly off by 1 more by first randomly sampling $\lfloor p_1 \times n \rfloor$ indices (without replacement) from our ID values. The dates corresponding to the sampled ID values are then altered according to the following. If the date is 31st December, then it is changed to 1st January of the next year. If the date is the 31st of any other month then it is changed to the 1st of the next month. For any other date the day value is simply increased by 1. An example of this process is given in Figure 4.11.

ID	dd	mm	yyyy
1	05	03	2001
2	07	03	2001
3	12	03	2001
4	13	03	2001
5	21	03	2001
6	31	03	2001
7	31	12	2001

$\longrightarrow$

ID	dd	mm	yyyy
1	06	03	2001
2	07	03	2001
3	13	03	2001
4	13	03	2001
5	21	03	2001
6	01	04	2001
7	01	01	2002

Random Sample of ID's: 1, 3, 6, 7

Figure 4.11: Example of simulating days randomly off by 1 more.

Problem 6: Months randomly off by 1 less

Given n dates and a proportion p_1, we simulate months randomly off by 1 less by first randomly sampling $\lfloor p_1 \times n \rfloor$ indices (without replacement) from our ID values. The dates corresponding to the sampled ID values are then altered by changing the month to the previous one. Note that this simulation does not alter the day or year values. An example of this process is given in Figure 4.12.

ID	dd	mm	yyyy
1	01	01	2001
2	17	02	2001
3	05	03	2001
4	27	03	2001
5	12	04	2001

$\longrightarrow$

ID	dd	mm	yyyy
1	01	12	2001
2	17	01	2001
3	05	02	2001
4	27	03	2001
5	12	03	2001

Random Sample of ID's: 1, 2, 3, 5

Figure 4.12: Example of simulating months randomly off by 1 less.

Problem 7: Months randomly off by 1 more

Given n dates and a proportion p_1, we simulate months randomly off by 1 more by first randomly sampling $\lfloor p_1 \times n \rfloor$ indices (without replacement) from our ID values. The dates corresponding to the sampled ID values are then altered by changing the month to the next one. Note that this simulation does not alter the day or year values. An example of this process is given in Figure 4.13.

ID	dd	mm	yyyy
1	07	03	2001
2	13	03	2001
3	21	03	2001
4	17	09	2001
5	15	12	2001

$\longrightarrow$

ID	dd	mm	yyyy
1	07	04	2001
2	13	03	2001
3	21	03	2001
4	17	10	2001
5	15	01	2001

Random Sample of ID's: 1, 4, 5

Figure 4.13: Example of simulating months randomly off by 1 more.

Problem: Inclusion of non-entry dates

In order to include non-entry dates it is necessary to create a different true date set. This is because the optimisation program should change the dates to be the same as the previous date. In order for this to work we first randomly sample $\lfloor p_1 \times n \rfloor$ indices (without replacement) from our ID values. We then duplicate the dates corresponding to these ID values in our true date set, and renumber our IDs. We then change the duplicated dates to a random date from 1st January 2001 to 31st December 2001. An example of this process is given in Figure 4.14.

ID	dd	mm	yyyy
1	05	03	2001
2	07	03	2001
3	12	03	2001
4	13	03	2001
5	21	03	2001

Random Sample of ID's:
1, 3, 4

↓

ID	dd	mm	yyyy
1	05	03	2001
2	05	03	2001
3	07	03	2001
4	12	03	2001
5	12	03	2001
6	13	03	2001
7	13	03	2001
8	21	03	2001

⟶

ID	dd	mm	yyyy
1	05	03	2001
2	15	08	2001
3	07	03	2001
4	12	03	2001
5	03	01	2001
6	13	03	2001
7	31	09	2001
8	21	03	2001

Figure 4.14: Example of simulating non-entry dates.

Choice of Simulation Parameters

In order to cover a wide variety of scenarios for each problem we used $p_1 \in \{0.25, 0.5, 0.75, 1\}$. For Problem 3 values of $p_2 \in \{0.5, 0.75, 1\}$ were used, but only when $p_2 > p_1$.

4.3.3 Optimisation of Simulated Data

The simulated data is now optimised according to the program given in Section 4.2, using varying values of α, β, γ, δ and ω. To ensure that simulations ran within a reasonable amount of time a 90 second timeout condition was introduced for the optimisation.

4.3.4 Simulation Results

Each set of simulated data was initially run for all parameter combinations where four of the parameters were set to 1, and the fifth parameter had values in $\{1, 0.25, 0.5, 0.75, 25, 50, 75, 100\}$. This was done in order to cover a wide variety of possible parameters. For each set of simulation parameters we then considered a 5-number summary for both accuracy and distance for the 100 simulations with that parameter set. We define accuracy as the proportion of optimised dates which are the same as the corresponding true date and distance as the average number of days the optimised dates are away from the corresponding true dates. Ideally, we are looking for optimisation parameters which give an accuracy close to one, and distance close to zero. We also considered whether any of the 100 simulations timed out. The results of these simulations can be found in the Github repository:

`https://github.com/AshleyDennisHenderson/Analysing-Australian-WW1-Diaries/tree/master/Simulation-Results`

It was found that the best results over the simulations for problems 1 - 7 was when $(\alpha, \beta, \gamma, \delta, \omega) = (25, 1, 1, 1, 1)$ or $(50, 1, 1, 1, 1)$ and the worst results were for $(\alpha, \beta, \gamma, \delta, \omega) = (1, 1, 1, 25, 1)$, $(1, 1, 1, 50, 1), (1, 1, 1, 75, 1), (1, 1, 1, 100, 1)$. For problem 8 it was found that the best parameter set varied depending on p_{true}, p_1 and p_2, with many parameter sets giving similar results. However, for problem 8 it was found that parameter sets in the form $(1, x, 1, 1, 1)$ or $(1, 1, 1, x, 1)$ where $x \in \{25, 50, 75, 100\}$ always gave the worst results. We also note that none of the simulations hit the 90 second time out condition.

4.4 Application of Date Extraction Process on World War I Diaries

This process of extracting dates and correcting them can now be applied to the World War I diaries. In Section 4.3 we found that for the first seven possible problems the best optimisation parameters were $(25, 1, 1, 1, 1)$ or $(50, 1, 1, 1, 1)$, whilst for the last problem the best parameters depended on the data. As we had no information on the best parameters for combinations of these problems we tried a range of parameters and then used the parameter set which gave the most accurate results.

Two different methods for determining the accuracy of the process, and hence the best parameter set, were considered. First, dates from a subset of the diaries could be manually extracted and compared to the optimisation results. However, this was not deemed feasible given the time constraints of this project. This was determined by manually extracting the dates from one of the diaries. This process took approximately 3.5 hours, and the diary had 11,660 words including 267 dates. From Table 3.4 we know that this document had less than the median number of words per diary. Therefore, given the amount of time to manually extract dates from this diary, and that we would need to do this for several diaries to get a reasonable representation of then entire data set, this method was not feasible. The second method was to compare the dates extracted from the diaries to the optimisation results and the known end dates of the diaries. Note, this method cannot confirm that the regular expressions are extracting all dates. However, it can help determine the best optimisation parameters for this data set given the extracted dates. In this method we consider the proportion of duplicated dates, and whether all the dates for the diaries are before their known end date. During the optimisation process we remove any duplicate dates. These duplicate dates should come from non-entry dates being changed to the same as the previous date or if an author has continued the entry on another page and written something such as "30th May cont.". However, a poor choice of parameters can lead to majority of the dates being changed to the same date. Therefore, we wish to have a low proportion of duplicated dates. If there is not enough information regarding the month or year within the diary it is possible for the optimisation program to drag dates out. This then gives diaries which go many years past their known end date. Ideally we would like no diaries to be after their known end date, and if they do we would like the average number of days a

diary is past its end date to be small.

To begin with, we tried the parameter set (25,1,1,1,1) as this was the best choice for problems 1-7. This had a low median duplication proportion. However, it was found that 245 diaries had dates past their end date, with the median average distance of these diaries past their end date being 177 days. This is not ideal, and we would like to reduce the number of diaries with dates past their end date. There are two parameters which could affect the year, γ and δ. Increasing γ would keep the optimised dates closer to the extracted years however this does not work due to the number of dates with missing years. Increasing δ will keep the dates closer together. However, as seen in the simulations, this can lead to less accurate results as it tends to force many dates to be the same, leading to a large number of duplicate dates. Hence, we need to find the balance between α and δ which works best for this data set. We tested a variety of parameter sets holding β, γ and ω constant at 1, and varying α and δ in the set $\{1, 10, 20, 25, 30, 40, 50\}$. Graphs of the median proportion of duplicated dates for each diary are given in Figure 4.15. Graphs showing the number of diaries with dates past their end dates and the median average number of days these diaries go past their end dates are given in Figures 4.16 and 4.17. From these graphs we find that large values of α, combined with small values of δ give results with the least amount of duplicates. However, large values of δ and small values of α give the least number of diaries with dates past their end dates and high values of both α and δ give diaries where the average number of days past the end date is small. Overall, there is not a parameter set which gives the best results in terms of all three of these, but the best compromise is setting both α and δ to 25.

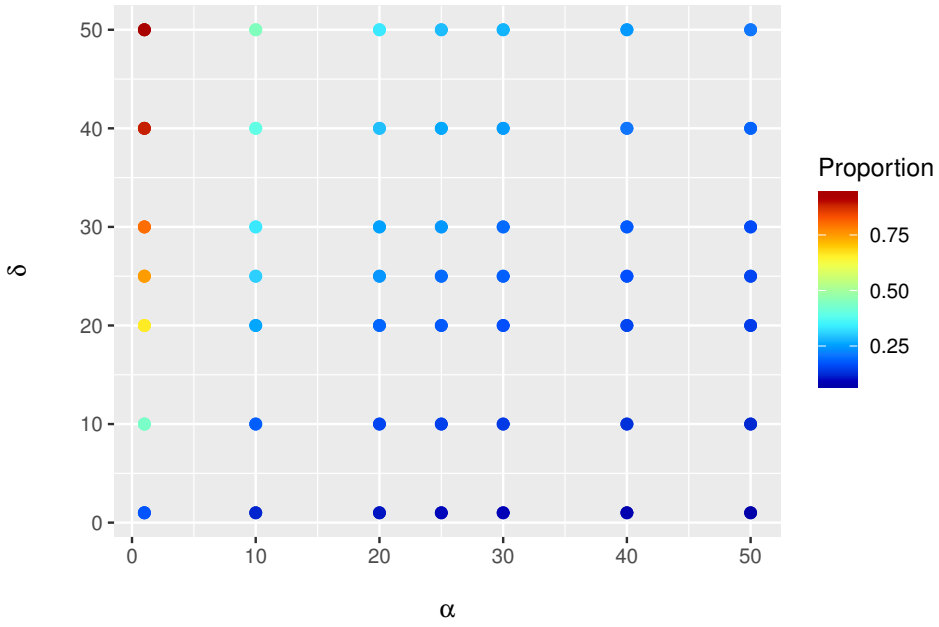


Figure 4.15: Graph showing the median proportion of duplicated dates for varying values of α and δ.

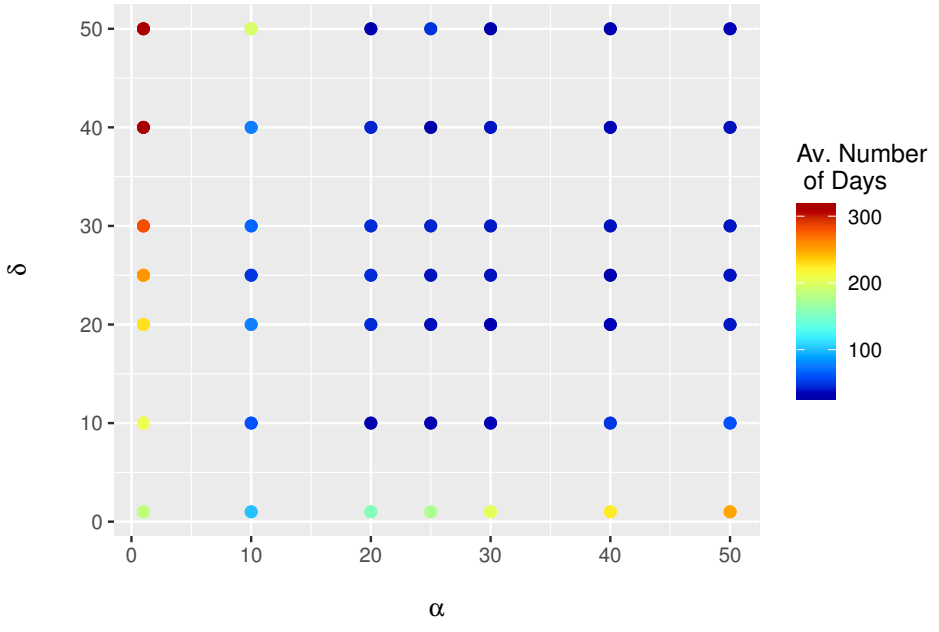


Figure 4.16: Graph showing the median average number of days past the end date for given varying values of α and δ.

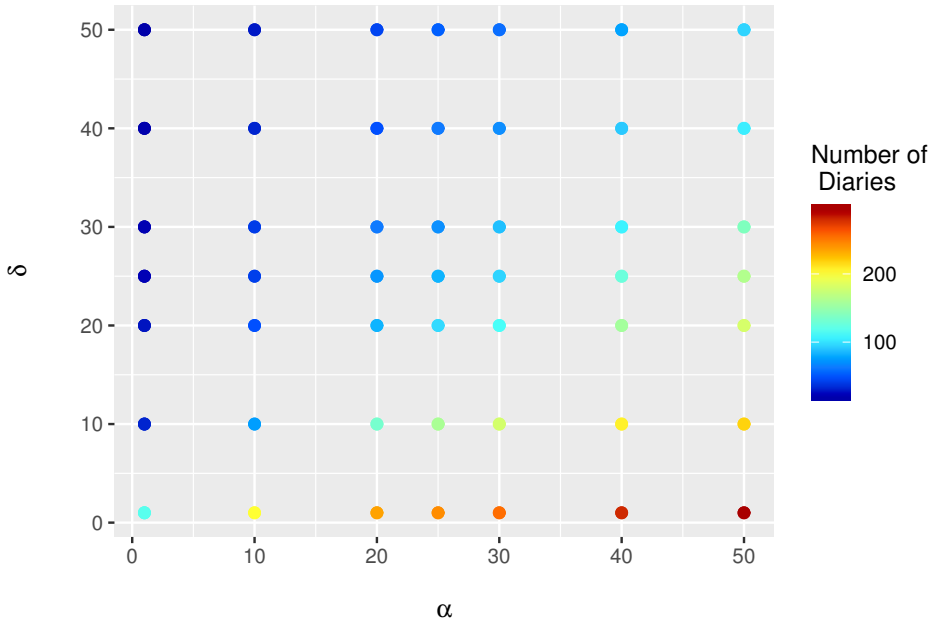


Figure 4.17: Graph showing the number of diaries past their known end date for varying values of α and δ.

The overall date extraction process is not completely accurate for our data set for several reasons, relating back to two main things. First, diarists wrote their diaries in many different styles and formats, and as such the way dates are included within diaries is widely varied. This is further complicated as documents are transcribed exactly as is from first page to last, and on each page left to right, top to bottom.

By considering all possible combinations made of day, month, year and DOW, as well as dates written in forms such as dd-mm-yyyy, we have been able to extract many dates. However, some diarists have not written dates in these forms. For instance, some diarists have simply used numbers (with or without a suffix), with no DOW, month or year, as a date. Likewise, some diarists only wrote the day of the week. If we were to extract text where just a day or DOW is used then we will also be extracting text which is not a date, for instance, when they refer to the 1st battalion. This would lead to less accurate results. It is also possible that diarists have used different spacing or punctuation within the date to what we have included.

Some diarists have also included summaries at the end of their diaries, listing dates for things such as pay received, letters sent or received, or their locations. For ex-

ample, Donald MacDonald's diary includes a list of letters he posted, as seen in Figure 4.18. This can lead to less accurate results as the program will believe these are dates for the next year.

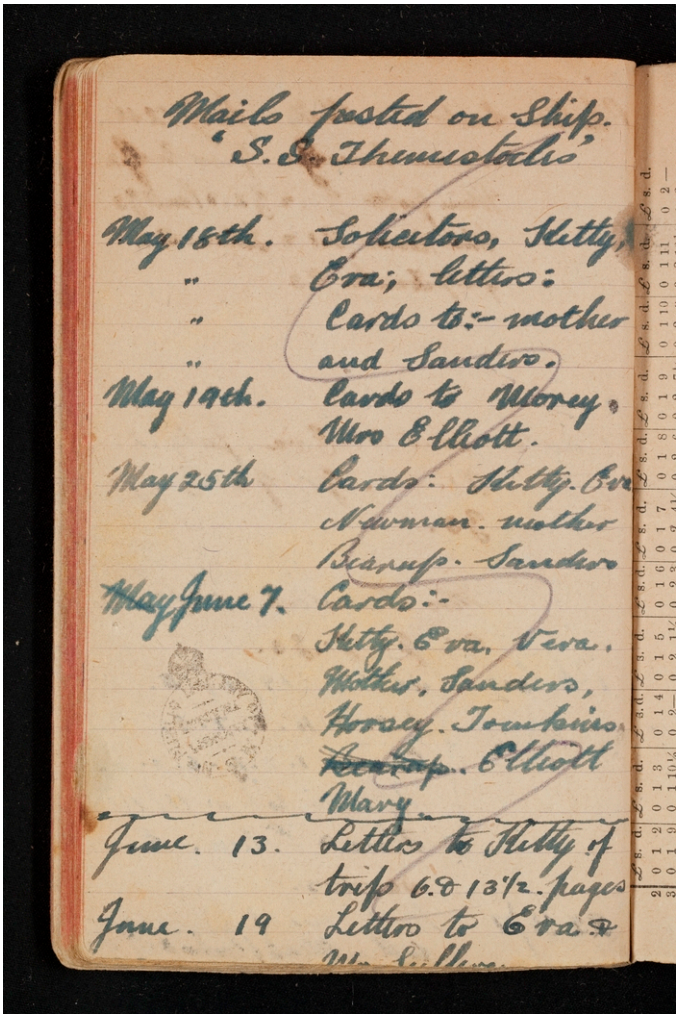


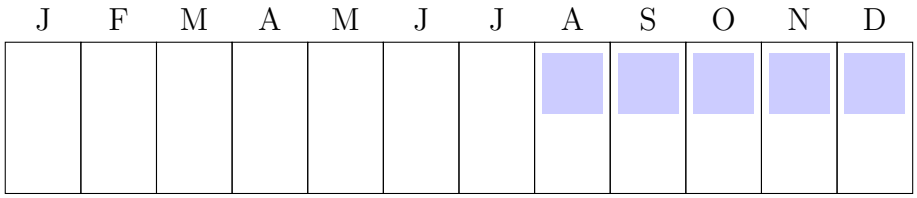
Figure 4.18: Diary page at the end of Donald MacDonald's diary showing a list of dates when letters were posted.

Finally, throughout these (and future) cleaning steps it was noticed that three diaries had dates which were out of order. The primary reason for this is that some diarists bought a diary with printed dates mid-way through the year. After finishing writing the entries for that year they then started writing the next year's entries at the start of that diary until they either reached the point they started the diary at, or until the end of the year. This is illustrated in Figure 4.19. There are several

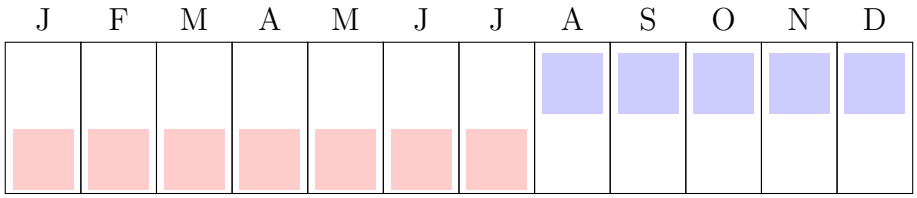
reasons diarists might have chosen to do this including not being able to purchase a new diary (for instance if they are on the front line), or not wanting to waste paper or money.

For example, Florence Holloway (a nurse), started her diary in September 1917 then continued writing entries until the end of October 1918. This meant that some pages included both a 1917 and 1918 entry as seen in Figure 4.20. Similarly, Arthur Freebody wrote his diary from March 1916 to March 1917. Two pages of his diary can be seen in Figure 4.21, showing where he has written the correct date next to the printed one.

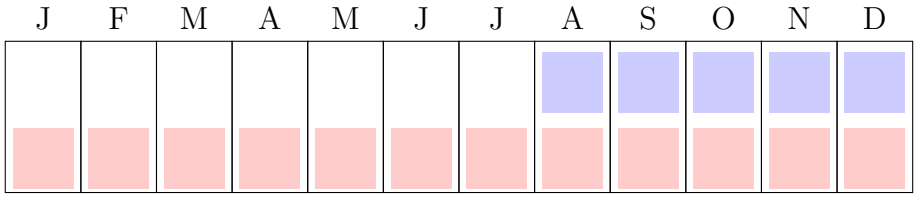
Alternatively, William Middleton's diary was out of order as the entries appear to have been written on a large piece of paper in segments which were out of order and facing different directions. When transcribing this the library has taken each segment as a page and ordered it such that the segments went from left to right, top to bottom. This can be seen in Figure 4.22.



The diarist begins by writing the entries for all the 1916 dates after purchasing the diary.



The diarist then writes their 1917 entries from the beginning of the diary.



Some diarists then continued writing their 1917 entries underneath the 1916 entries until the year finished.

Legend: 1916 entries, 1917 entries

Figure 4.19: Illustration of diaries with printed dates where the diarist has started writing entries mid-year then written the next years entries at the start of the diary. The years 1916 and 1917 have been used as an example.

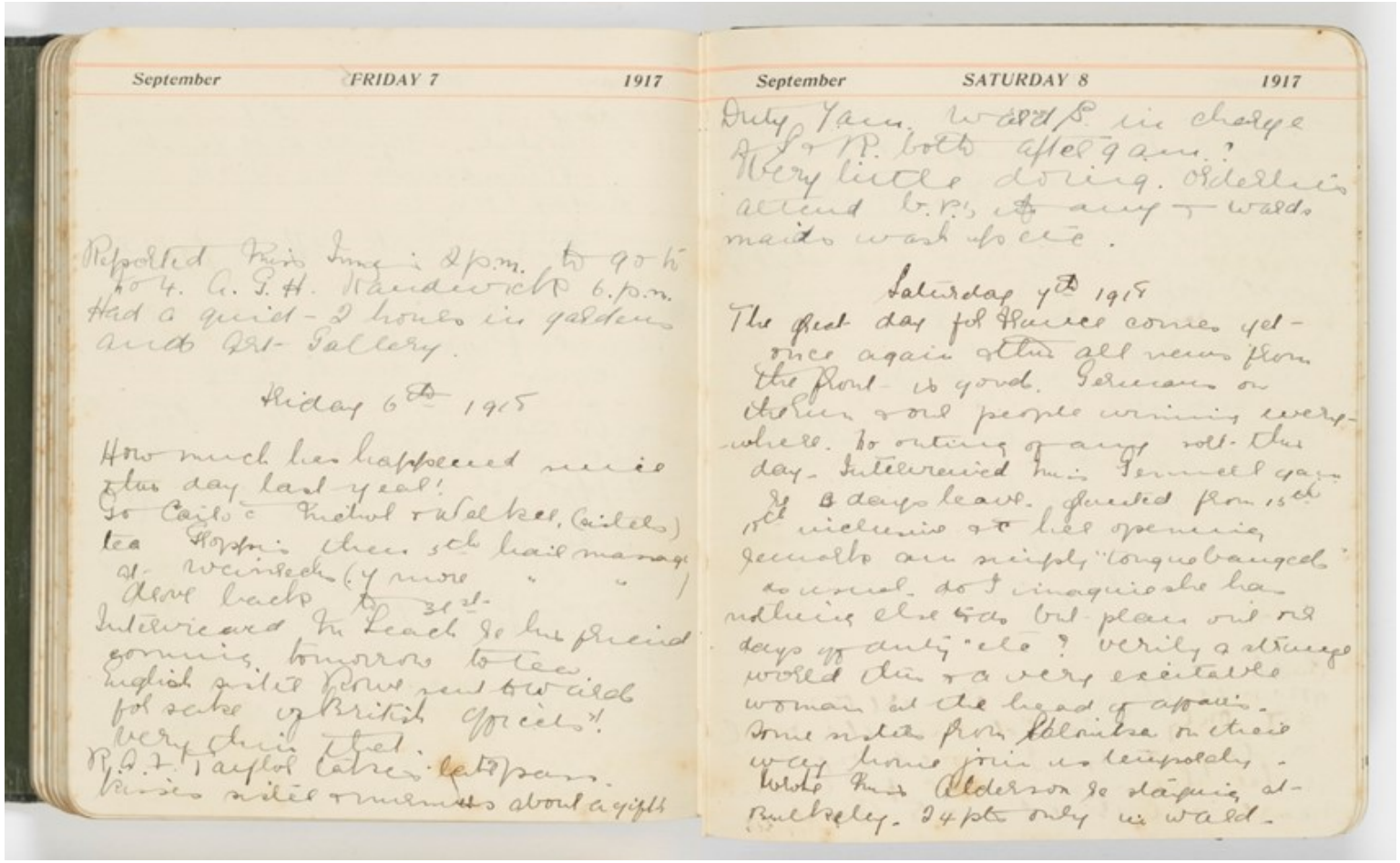


Figure 4.20: Florence Holloway's diary showing 1917 entries with 1918 entries below them.

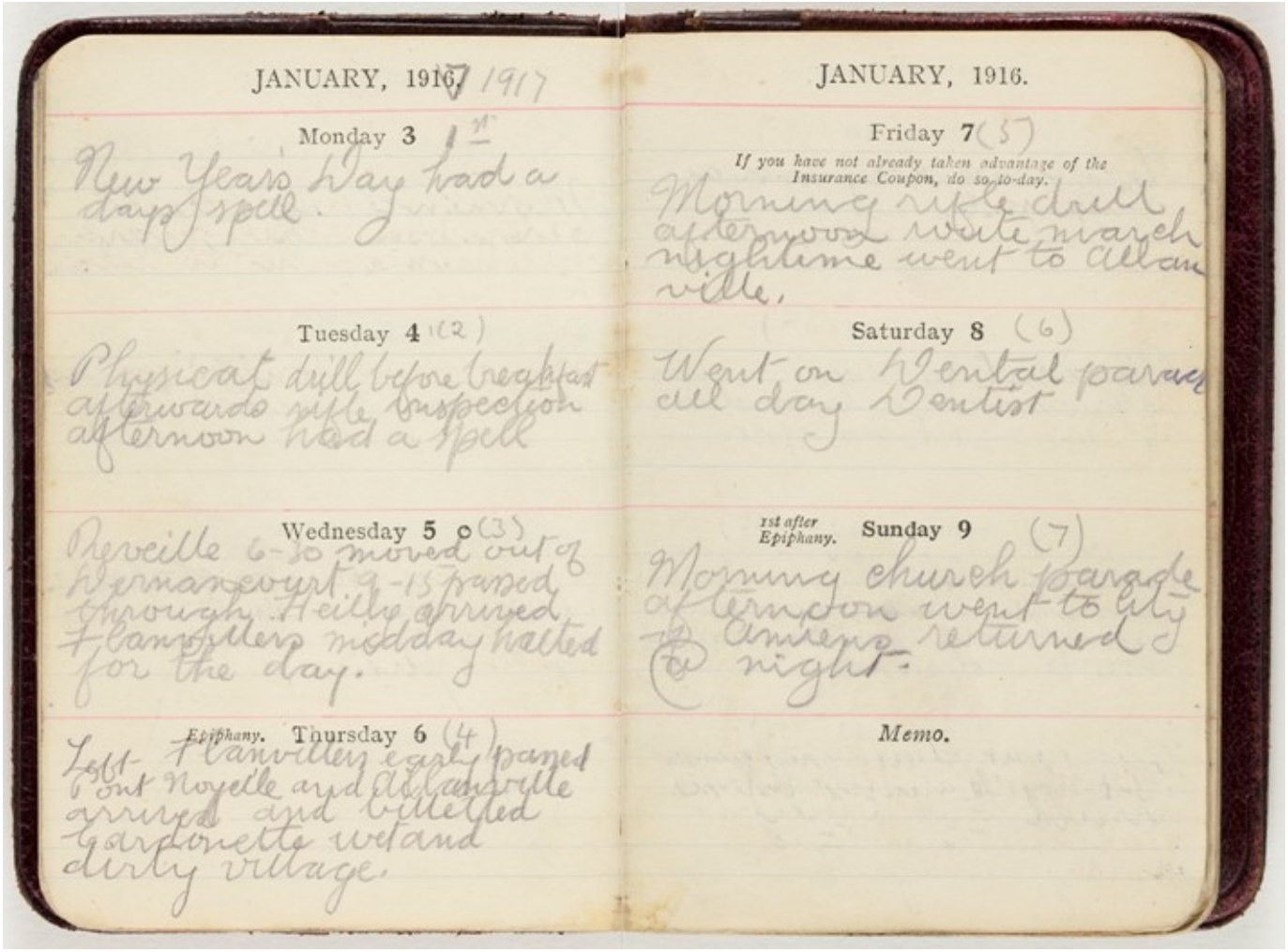
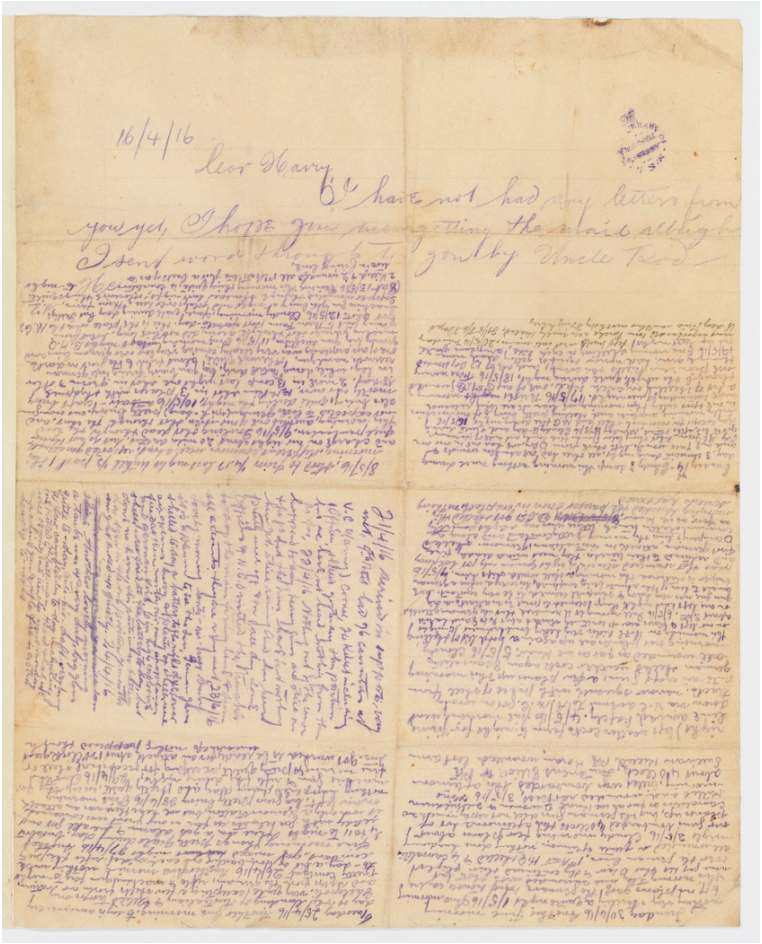


Figure 4.21: Arthur Freebody's diary that went from March 1916 to March 1917, showing where he has altered a 1916 diary for 1917 by putting the correct day to the right of the printed date.



Letter: 16/04/16 1 1	
08/05/16 2 to 13/05/16 6	14/05/16 3 to 21/05/16 7
21/04/16 4 to 24/04/16 2	04/05/16 5 to 07/05/16 5
25/04/16 6 to 29/04/16 3	30/04/16 7 to 03/05/16 4

Figure 4.22: First Page of William Middleton's diary that went from March to May 1916. Note that the diary entries were written in 6 segments of the paper, facing different directions and in no particular order in terms of date. To the right of the image is a diagram of the page showing the start and end date of each segment, the order it was transcribed in (in red) and the order it was written in (in blue).

One way we could improve the date extraction process is by including a condition in our optimisation program to keep all dates before a known end date. This was not included in the original program as we wanted a program that was more accessible and it is not always necessarily easy to get the end date for documents. As seen in these diaries, some diarists have summaries at the end, also finding an end date is harder as you have to search back through text to find it, whereas a start date is usually closer to the first few lines. However, since we have known end dates for these diaries, we tried including this to see if it improved our accuracy.

Since we are forcing all dates to be before the known end date we can no longer look at how far diaries go after their end date as a measure of accuracy. Instead we consider the proportion of dates changed in the optimisation process as well as the proportion of duplicate dates. Also, as we are using the end date to control the year values in our optimised dates we no longer need to consider varying δ. So we hold all parameters at 1 except α which we vary in the set $\{1, 25, 50, 75, 100, 125, 150, 175, 200\}$. The results of these trials are given in Figure 4.23. It is seen that the proportion of duplicate dates decreases as α increases. However, after $\alpha = 25$ the proportion of dates changed in the process increases. Based on this, we conclude that the best parameter set is $(25, 1, 1, 1, 1)$. Overall, including a condition on the end date in the optimisation program has increased the accuracy of our dates. Consequently, to create our date/entry table we use this method with the parameter set $(25, 1, 1, 1, 1)$.

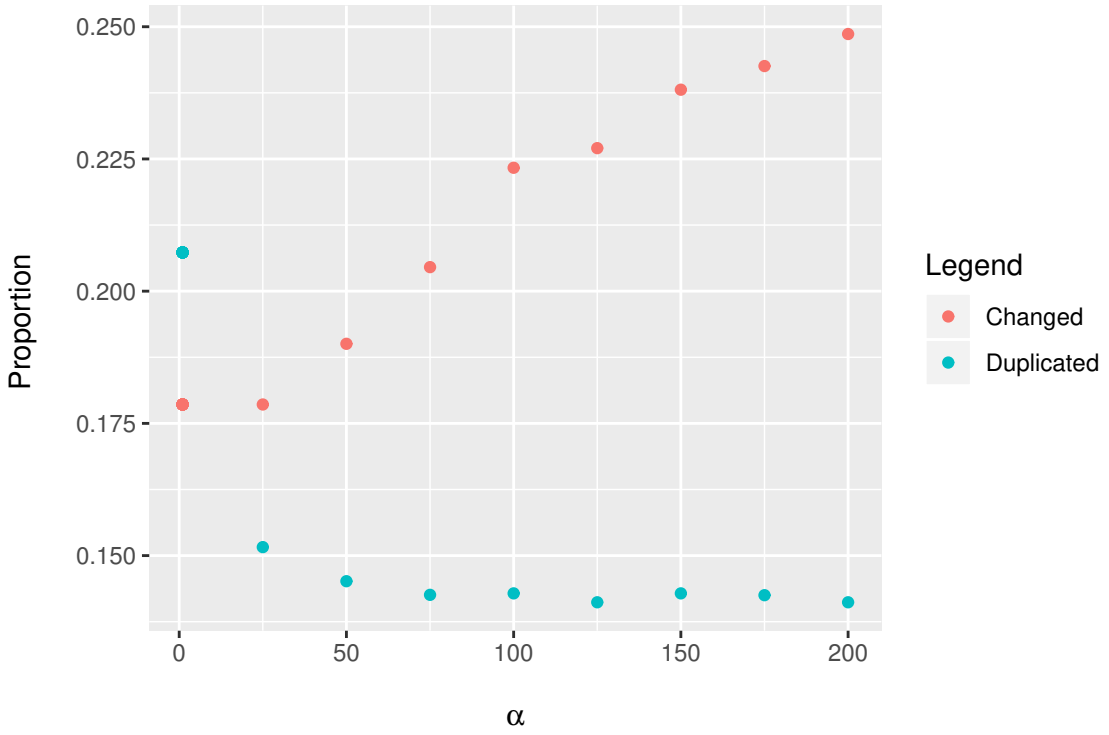


Figure 4.23: Graph showing the median proportion of changed dates and duplicate dates. It is seen that after $\alpha = 25$ the proportion of changed dates increases, so we conclude that $\alpha = 25$ gives the best results.

There are some things we can try to improve our date extraction method, including:

- Testing other methods or REs to improve the extraction phase of this process. This could include trying named entity recognition (NER), or looking into the structure of sentences to identify differences between the date a diary was written, and a date-like string within the text.
- Running simulations to determine the best parameter set when combinations of problems are present, as well as, for both when there is and is not an end date included in the program.

We can now look at how many words we have in our diary collection per month and this is shown in Figure 4.24. Note that the majority of the data is between August 1914 and December 1919.

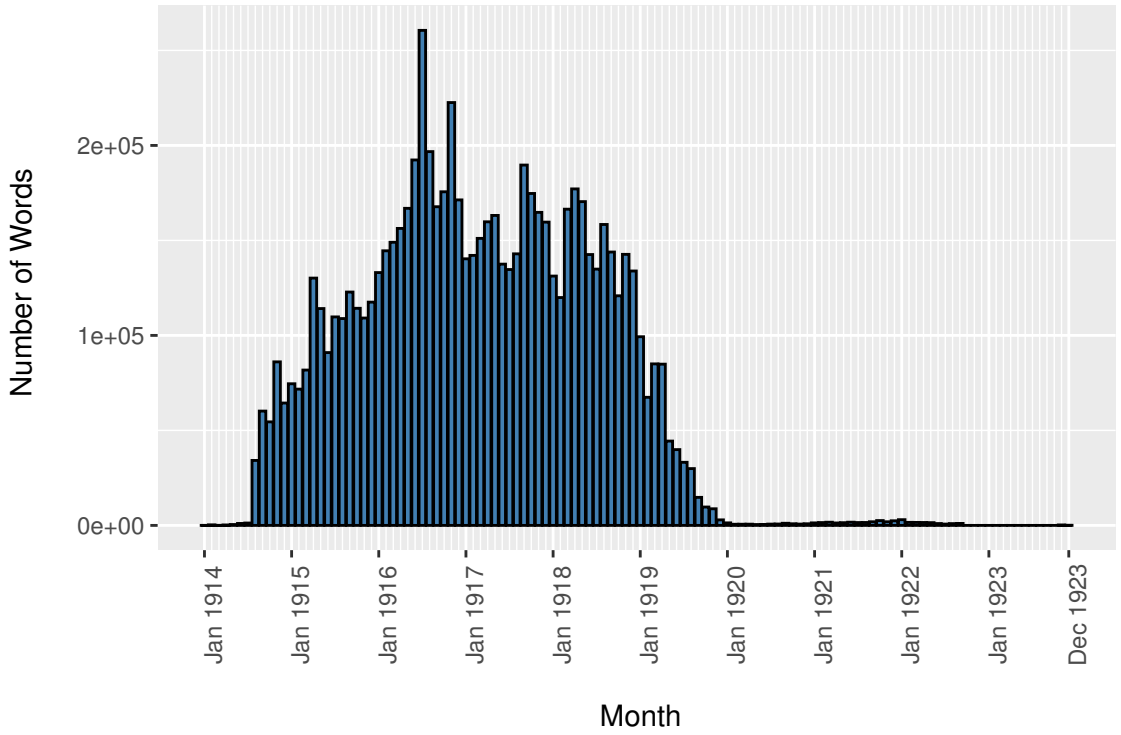


Figure 4.24: This graph shows the number of words written in our entire diary collection per month. It can be seen that the majority of entries are written between August 1914 and December 1919.

Now that our diaries are in a date/entry format we can analyse them to determine what the soldiers wrote about and how they felt about this. The next chapter focuses on determining what the soldiers wrote about by considering word frequencies, tf-idf, and topic modelling.

Chapter 5: Topic Analysis

Now that the data is cleaned and in a date/entry format we can use various techniques to determine what topics the diarists wrote about, and how these topics changed over time. These techniques include the consideration of word frequencies, tf-idf (term frequency - inverse document frequency) and topic modelling. The theory behind these techniques, and the results of our analysis, are discussed in Sections 5.1, 5.2 and 5.3, respectively. In Section 5.4 we discuss the similarities and differences between our techniques as well as possible future work.

5.1 Word Frequencies

The first step in understanding our data is to consider word frequencies, that is, how often each word is used. We use word frequencies to get a sense of the overall data set. Figure 5.1 gives a word cloud with the 100 most frequent words in the entire data set. We have manually labelled these words as being common, homonyms or war-related words as depicted by the colours in Figure 5.1. Common words are those that do not have a specific meaning regarding war. Homonyms are words that have multiple meanings, and in this case are words which have different meanings depending on whether they are put in the context of war or not. For instance, neither light nor horse are necessarily related to war. However, when placed in this context they would be referring to the Light Horse. War-related words are those which are related to fighting even without being specifically placed in that context. From Figure 5.1 we see words that are expected in diaries about World War I, such as *wounded*, *gun*, *camp*, etc. However, we also note that over half of the words in this graph are common words, and that the most frequent words, which are larger in size and in the centre of the graph, are primarily common words. This shows that even though they are fighting in a war, some of the most important things the diarists write about are common things such as time of day and meals.

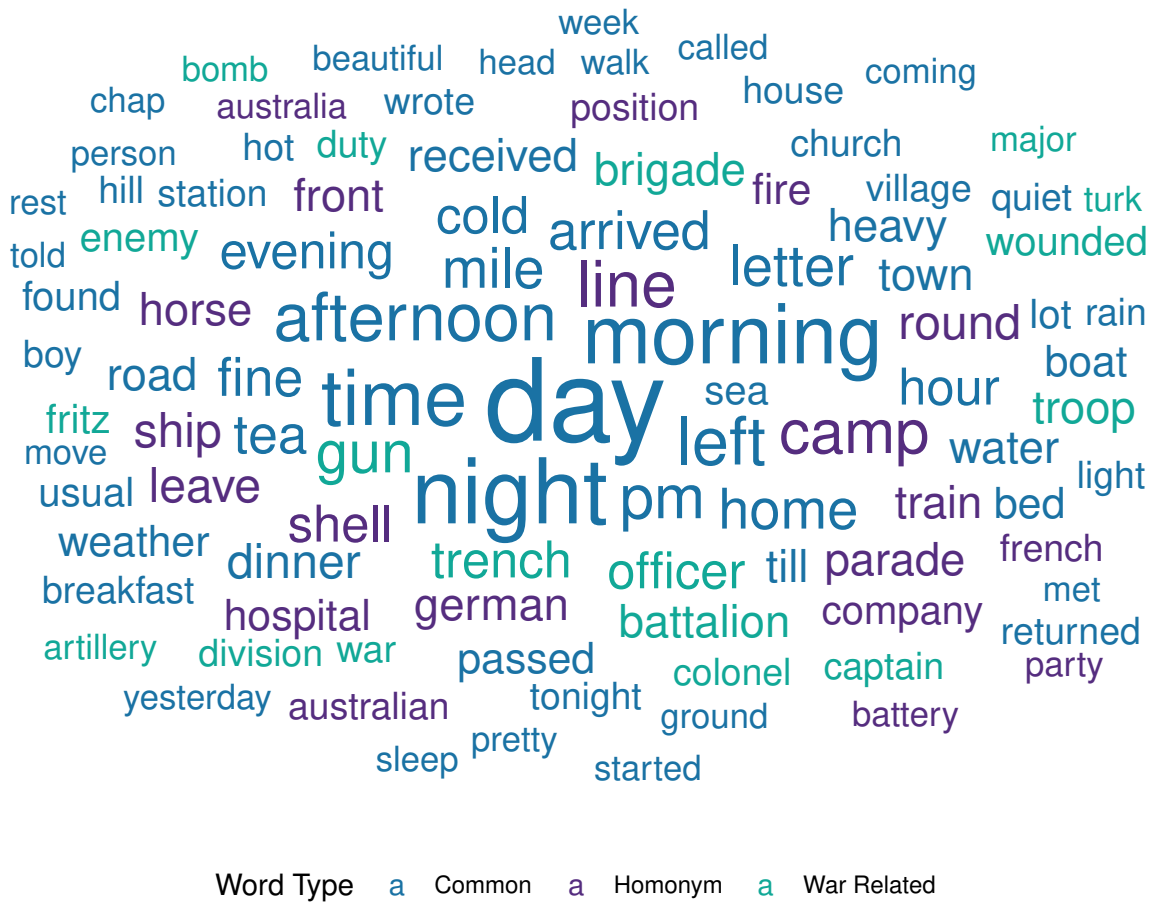


Figure 5.1: Word cloud of the 100 most frequent words in the entire World War I diaries data set. Note that the more frequent words are larger in size and words are coloured based on whether they are common, homonyms or war related.

We can also consider the frequency of particular n-grams to investigate subjects of interest, where n-grams are a string of n words. The subjects we chose to investigate were politics and health.

Throughout the war, politics was important due to the effects of policies on both troops and those left at home. Three important events were the 1916 and 1917 conscription referendums and the 1917 election. To investigate this we considered n-grams regarding Prime Minister Billy Hughes, voting and conscription, with a full list of the n-grams given in Table 5.1. The frequencies of each n-gram set is shown in Figure 5.2.

	n-grams
Billy Hughes	billy hughes, mr hughes, wm hughes, w m hughes
	billie hughes, william hughes, prime minister
Vote	vote, voting
Conscription	conscription

Table 5.1: List of n-grams used to investigate politics throughout the war.

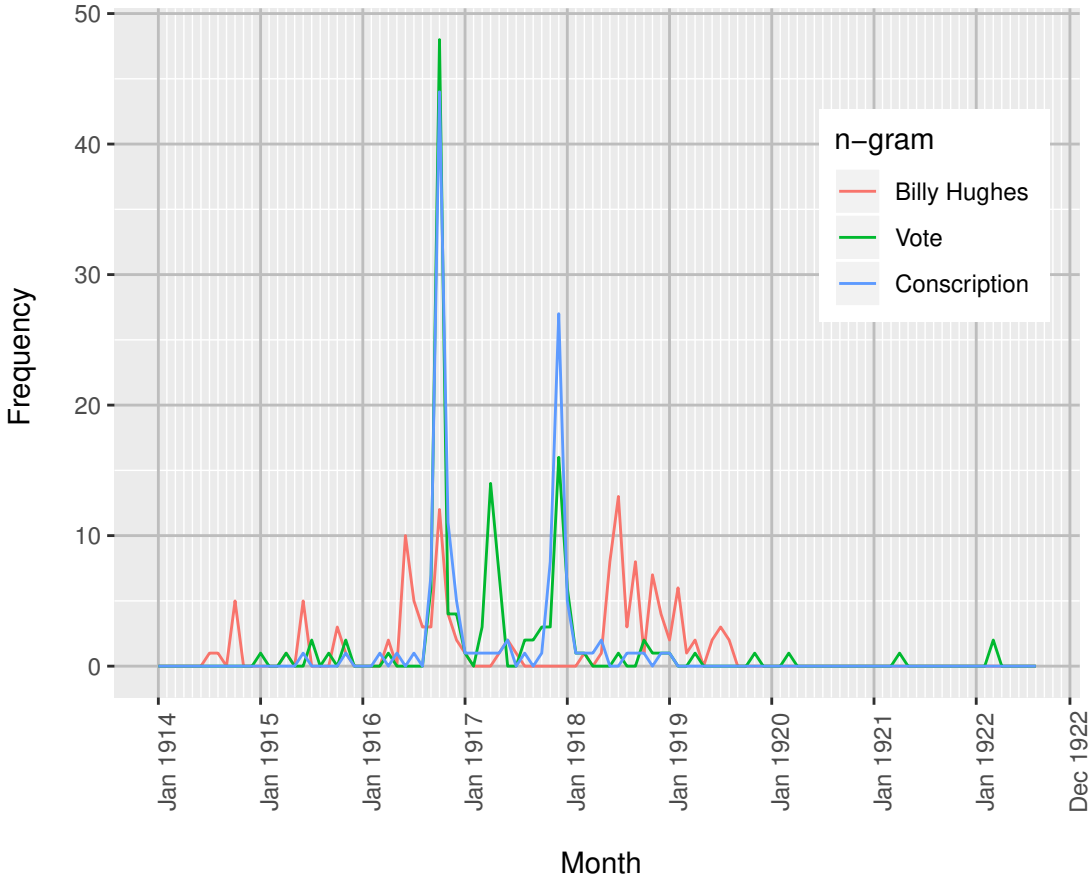


Figure 5.2: Frequencies of the politics n-grams from Table 5.1 over time.

In Figure 5.2 we see two main peaks for conscription in October 1916 and December 1917, corresponding to the two conscription referendums [37]. We also see three peaks for the n-grams regarding voting, in October 1916, April 1917, and December 1917. These correspond to the two conscription referendums as well as the Australian election in May 1917 [38]. This shows that these political events were

important to those already on the front line. Interestingly, whilst the Prime Minister is mentioned throughout the course of the war, he is not mentioned during either the 1917 election or 1917 conscription referendum. Through close reading of the diaries we see that many references to the Prime Minister are when he is visiting and inspecting troops. For instance, in Robert Harris' war diary:

> "Mr Wm Hughes Prime Minister of Australia addressed the troops, but he failed to raise much enthusiasm as his arrival had been delayed & all by the time we marched off all estaminets had closed.".

Next we considered two different aspects to do with the subject of health. The first are words related to being sick, and for this we consider n-grams to do with the word sick, influenza and more specifically Spanish influenza. The second aspect is medical staff and for this we consider doctors, nurses, dentists, and stretcher bearers. The n-grams used to search for both aspects of health can be found in Tables 5.2 and 5.3, respectively. The frequencies of each n-gram set can be seen in Figures 5.3 and 5.4, respectively.

	n-grams
Flu	influenza, flu, grippe
Spanish Flu	spanish flu, spanish influenza, spanish grippe
Sick	sick, sickness, ill, illness

Table 5.2: List of n-grams used to investigate sickness and influenza throughout the war.

	n-grams
Doctor	doctor, physician, doc
Nurse	nurse
Dentist	dentist
Stretcher Bearer	stretcher bearer

Table 5.3: List of n-grams used to investigate medical workers throughout the war.

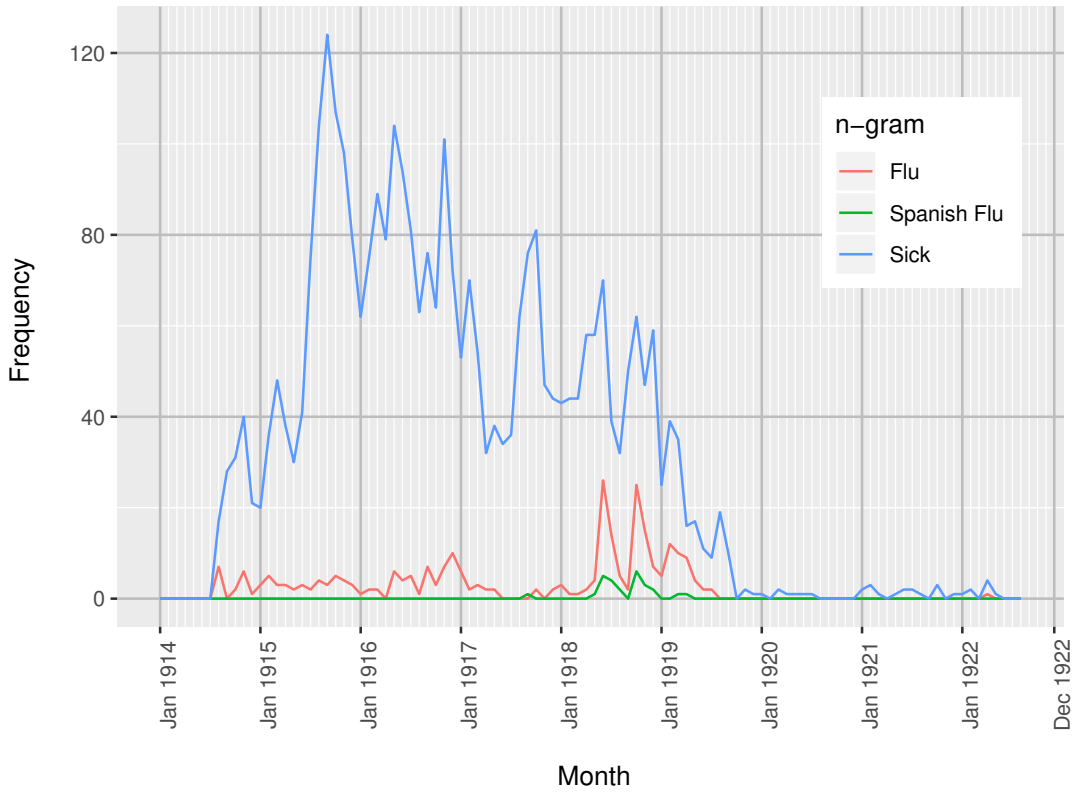


Figure 5.3: Frequencies of n-grams used to investigate sickness and influenza from Table 5.2 over time.

From Figure 5.3 we note that influenza is mentioned throughout the entire war, but has two peaks in mentions in June and October 1918, with a smaller peak in early 1919. These peaks correspond to the three waves of Spanish influenza [39]. We can also see two small peaks for Spanish influenza in June and October 1918 corresponding to the first two waves. However, these peaks are much smaller than those when just considering influenza, suggesting that at the time it was more commonly referred to as influenza rather than Spanish influenza. Through close reading of the diaries from this time period, we note that this is the case and that the Spanish flu impacted the soldiers greatly, with many falling ill and being evacuated, leave being cancelled, ships being quarantined, and soldiers dying. For example:

> "There is an epidemic of mild influenza throughout the continent. Spain was the first to experience it but now the sickness has spread through England, France & Germany. The quietness on the enemy front is at-

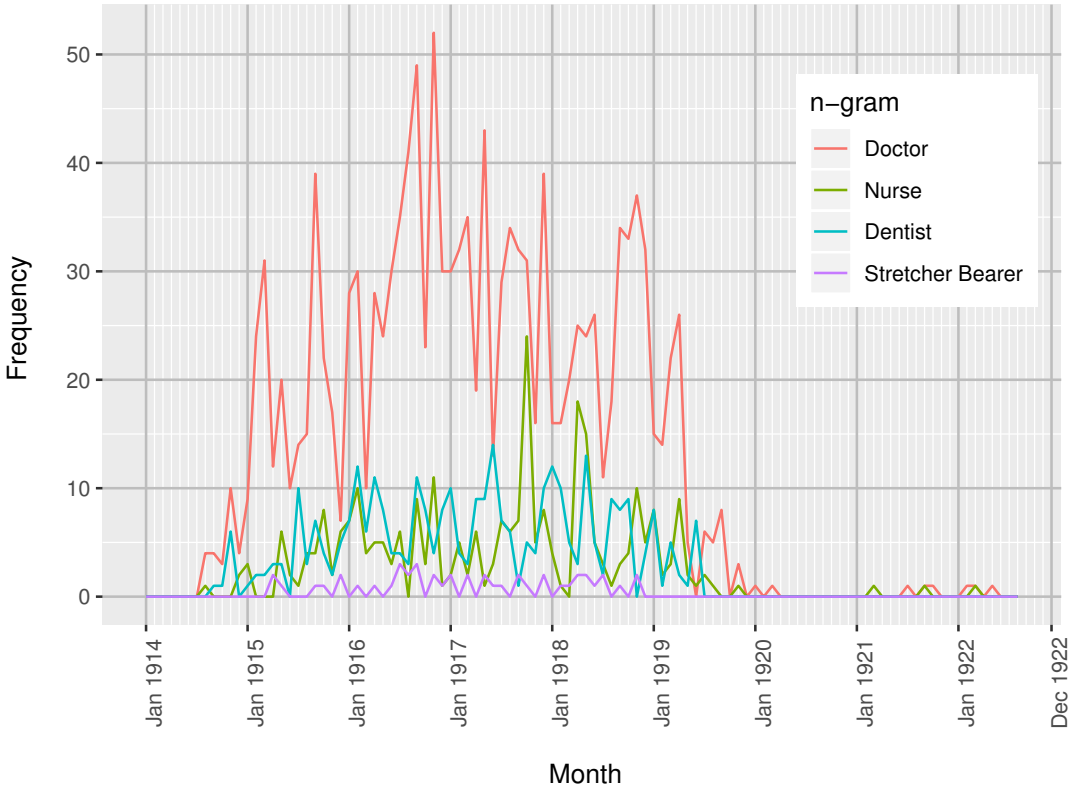


Figure 5.4: Frequencies of n-grams regarding medical workers from Table 5.3 over time.

tributed to this influenza. We are evacuating large numbers of our own men daily who are down with it.", Walter Rainsford

"Sickness has broken out all over Germany, It is called the Spanish Grippe. The people are dying in hundreds from it. The symptoms are much like Ordinary Influenza only it is much more severe.", Henry Parsons

"Influenza getting very prevalent and a large number of deaths.", Rudolph Cox

"The influenza epidemic now broke out amongst us and it was sad to see so many of our men dying with the 'flu ones who had gone all through the war without a scratch.", Verdi Schwinghammer

"Admiral Fergusson transferred his flag to the "Melbourne" as we were the only ship of our squadron able to go to sea as the "Dublin" was in

dock, and "Birmingham" and "Sydney were both in quarantine owing to Influenza Epidemic", George Iles

We also note that words regarding being sick are used frequently throughout the war. This is unsurprising given Australia had approximately 431,000 cases of sickness or non-battle related injuries throughout the war [20].

From Figure 5.4 we note that all medical staff were consistently mentioned throughout the war, until mid 1919, of whom doctors were the most prevalent. This is expected given the numbers of sick and wounded throughout the war. Interestingly, dentists were mentioned just as much as nurses. This was because if a soldier was unable to eat then he was unable to fight and dental disease, including broken teeth or dentures, was very common throughout the war [40].

5.2 Tf-idf (Term Frequency - Inverse Document Frequency)

Tf-idf (term frequency - inverse document frequency) gives us information about the most distinctive terms in a document. This is done by considering the frequency of a term in a specific document compared to the frequency of that term in all documents [41]. The classical formula for the tf-idf score of the ith term in the jth document is given by

$$\text{tf-idf}_{i,j} \quad = \quad \text{tf}_{i,j} \times \log\left(\frac{N}{\text{df}_i}\right),$$

where N is the total number of documents in the collection, $\text{tf}_{i,j}$ is the frequency of term i in document j and df_i is the number of documents term i appears in [42].

Terms which are used frequently in only a small subset of documents will have high tf-idf scores for those documents [41]. Conversely, terms which are not used frequently within a document or terms which are used frequently in all documents will have low tf-idf scores.

Tf-idf scores were calculated for our data using each year from 1914 - 1919 as a document, and all entries from 1920 - 1923 as another document. The years 1920 - 1923 were combined as one document as this was after the war finished, and, as seen in Figure 4.24, there is much less data for these years. We then considered the words with the highest 30 tf-idf scores for each document. These results are given in Figures 5.9 - 5.15 and are discussed below. Note that if words have the same tf-idf score they are given in alphabetical order.

For 1914 (Figure 5.9) the majority of words are locations or ship names, primarily associated with the Australian occupation of the German settlements in New Guinea. The locations mentioned are shown on the map in Figure 5.5. The Cocos Islands are also shown on this map as this is where the German raider Emden was sunk [43]. Note that "alexi" in Figure 5.9 is referring to the town *Alexishafen (Alexis Hafen)*. Also, "wilhelmshafen" is referring to the town *Friedrich Wilhelmshafen*, which was renamed *Madang*. This is known through close reading of the diaries:

> "Today I took an armed party up to Alexis Hafen which is the next harbour of any size north of Madang.", Clarence Read

> "However, as the whole of the German residents at Friedrich Wilhelmshafen surrendered and took the oath of neutrality, . . .", William Holmes

The ships mentioned in Figure 5.9 are:

- Aorangi
- Emden
- Ibuki
- Komet
- Lili
- Markomania
- Meklong
- Moresby
- Nusa
- Orvieto
- Siar

These are known to be ships based on close reading of the text:

> "We are to await a store ship "Aorangi" from Sydney", Vivian Little

> "A Dutch official, . . . requested that "Emden" and "Markomania" must, . . . leave the harbour as this was no place for coaling and and besides both steamers had already been more than 24 hours here.", Franz Bordeaux

> "the Japanese armoured cruiser "Ibuki" was also accompanying these warships to Albany having joined them en route", George Iles
>
> "The "Komet" was captured in a small harbour on the North-west Coast of New Britain, about 160 miles from Rabaul", William Holmes
>
> "It proved to be the "Lili", about 40 ft overall, & well adapted to our purpose. As is usual, however, in such cases something vital was missing. We therefore decided to sail her back.", Clarence Read
>
> "He was placed on board the "Meklong" and allowed no communication whatever with any outsiders.", William Holmes
>
> "Arrived at 'Tari at noon on Friday, and were surprised to see the "MORESBY" at anchor in the Harbour. ", Alan Fry
>
> ""Nusa" sailed at 9-0 p.m. for Carden Island", William Holmes
>
> "The "Orvieto" is the flagship. She is the first ship in our column", Cameron Robertson
>
> "Mr. Taifert, Manager of the New Guinea Company, was very anxious to know whether he could send word for the Steamer "Siar" to come to Rabaul.", William Holmes

The top 30 words for tf-idf in years 1915 - 1919 include various locations around Europe and the Middle East. All these locations are shown in Figures 5.7, and 5.8.

For 1915 (Figure 5.10) we find many words regarding combat, e.g. *gun*, *artillery*, and *wounded*, as well as *Turkish* and *landed*, showing that in 1915 soldiers were writing about the Gallipoli Campaign. Further, we see place names from this campaign including *Achi Baba*, *Quinn's Post*, *Gaba Tepe* and *Port Mudros* (on Lemnos Island), which are shown on the map in Figure 5.6. We also see the words *camp*, *Cairo* and *Alexandria*, showing we also have entries regarding the training of Australian soldiers in Egypt.

For 1916 - 1918 (Figure 5.11 - Figure 5.13) we once again see many words related to battle such as *trench*, *troop*, *bomb*, *artillery* and *battalion*. In each of them we also see the word *German* or the nicknames for German soldiers: *fritz* and *hun*, which is related to the fighting on the Western Front. Further, we see the word *wadi* in 1917/1918 and locations in the Egypt in 1916 which are related to the fighting in the Middle East. The primary differences between these years is that all the locations mentioned in 1916 are in the Middle East, whilst all locations mentioned in 1917/1918 are in Europe.

For 1919 (Figure 5.14) we no longer see words related to battle since the war is now over. Instead we see locations around Europe, the *Chemnitz* (a ship which brought soldiers back to Australia), and words regarding leisure activities.

Due to the limited amount of data for 1920 - 1923 (Figure 5.15), tf-idf analysis does not provide any meaningful results. We do see the word *dad* as well as locations in Australia, such as *Hurstville* (a suburb in Sydney), showing that the small amount of diary entries we have from this time are about being back home in Australia. However, due to the limited amount of data the results are very specific to what particular authors talk about, as well as their spelling mistakes.

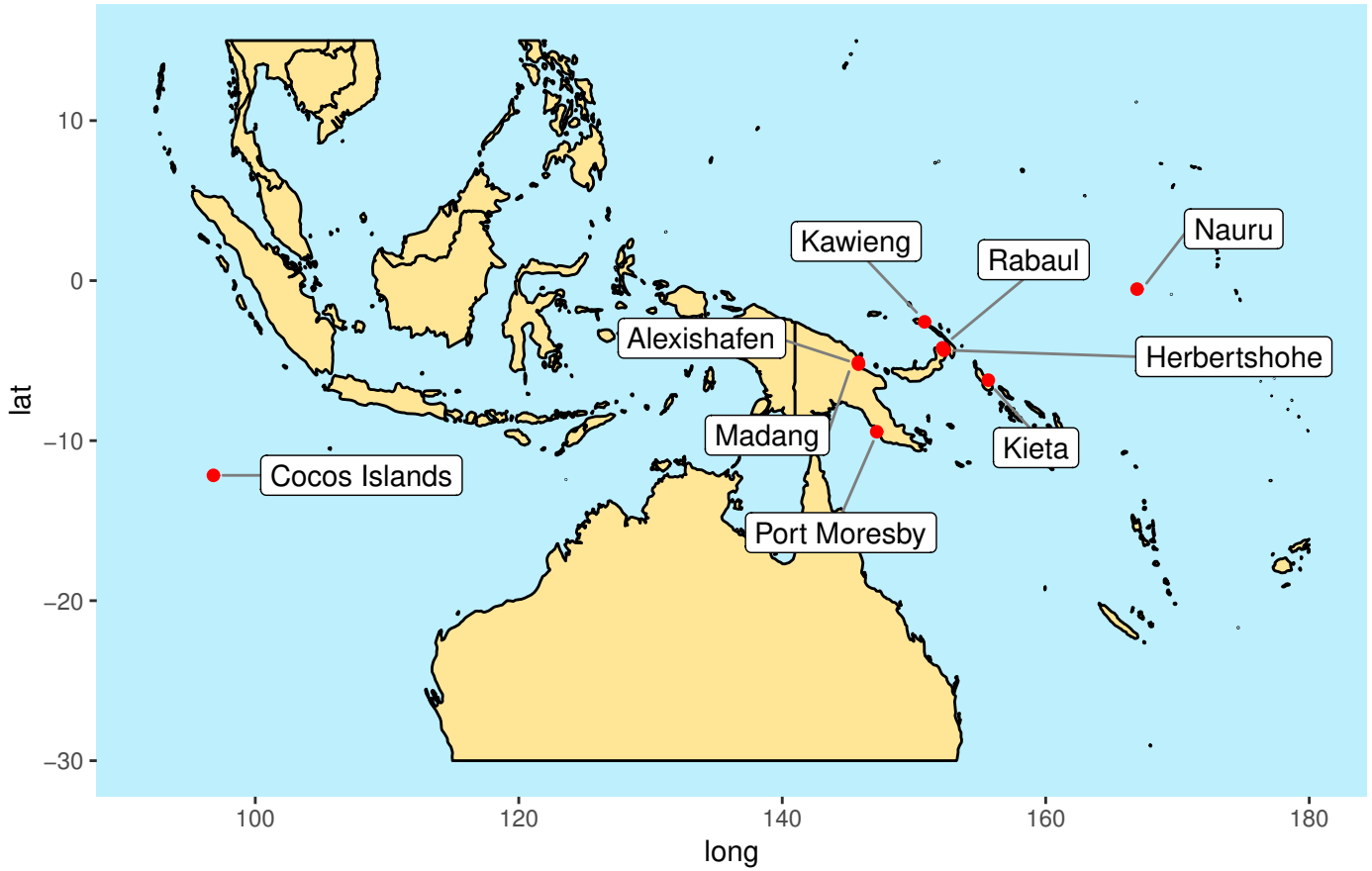


Figure 5.5: Locations in New Guinea seen in our tf-idf analysis for 1914. The location of the Cocos Islands (where the Emden was sunk) is also shown.

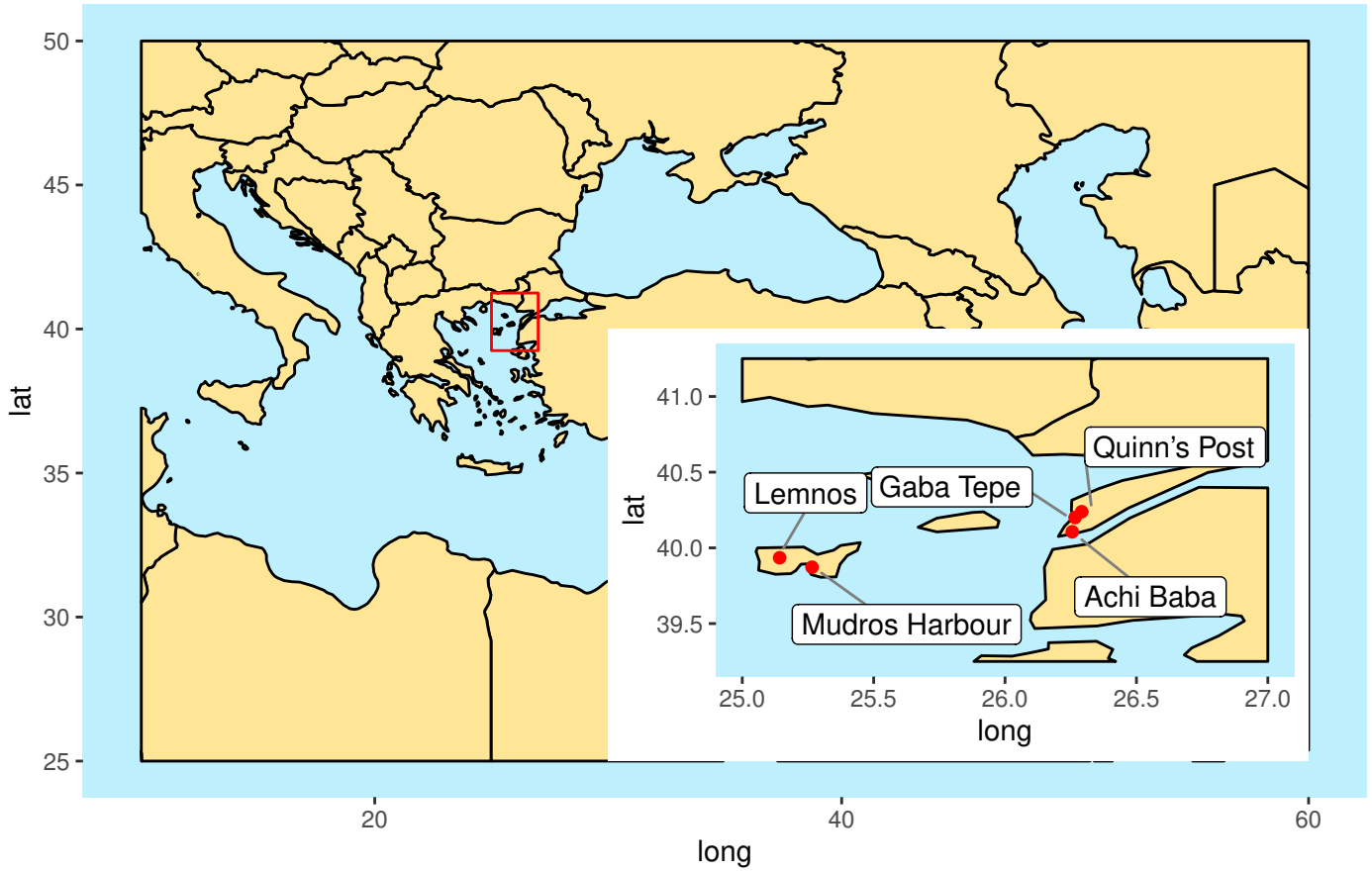


Figure 5.6: Locations around Gallipoli that are seen in our tf-idf analysis for 1915.

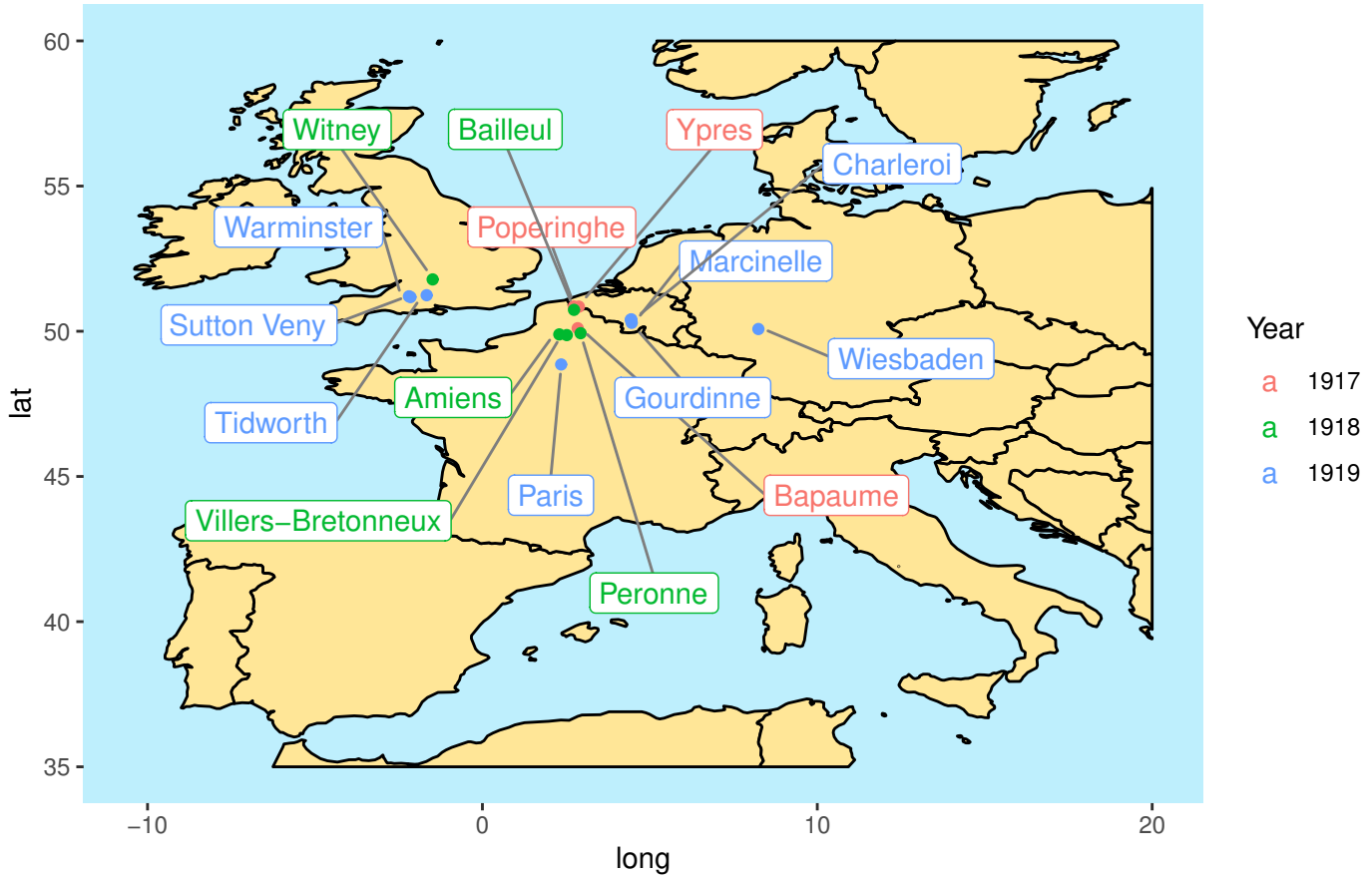


Figure 5.7: Locations in Europe seen in our tf-idf analysis for 1917 - 1919.

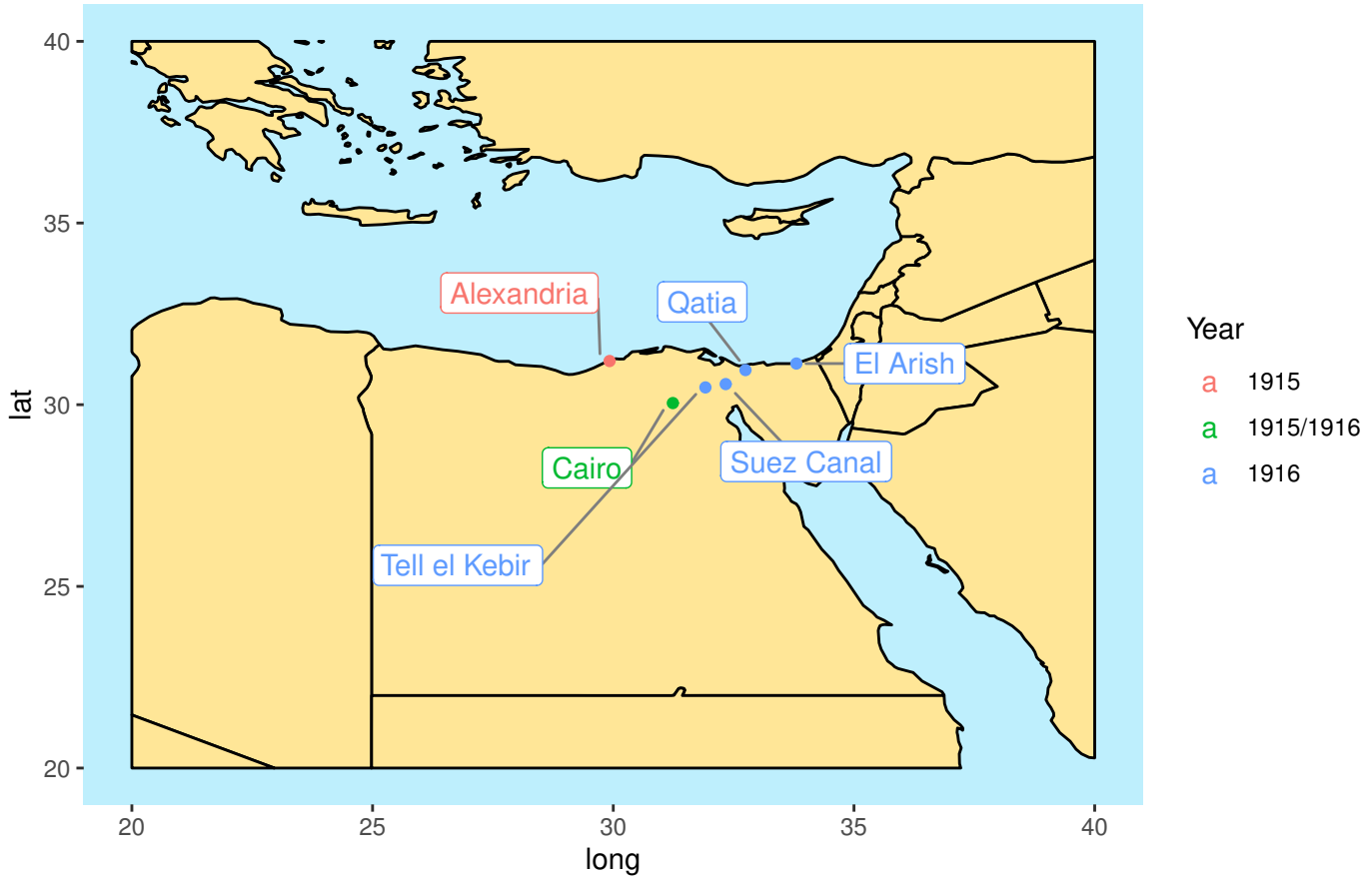


Figure 5.8: Locations in the Middle East seen in our tf-idf analysis for 1915 and 1916.

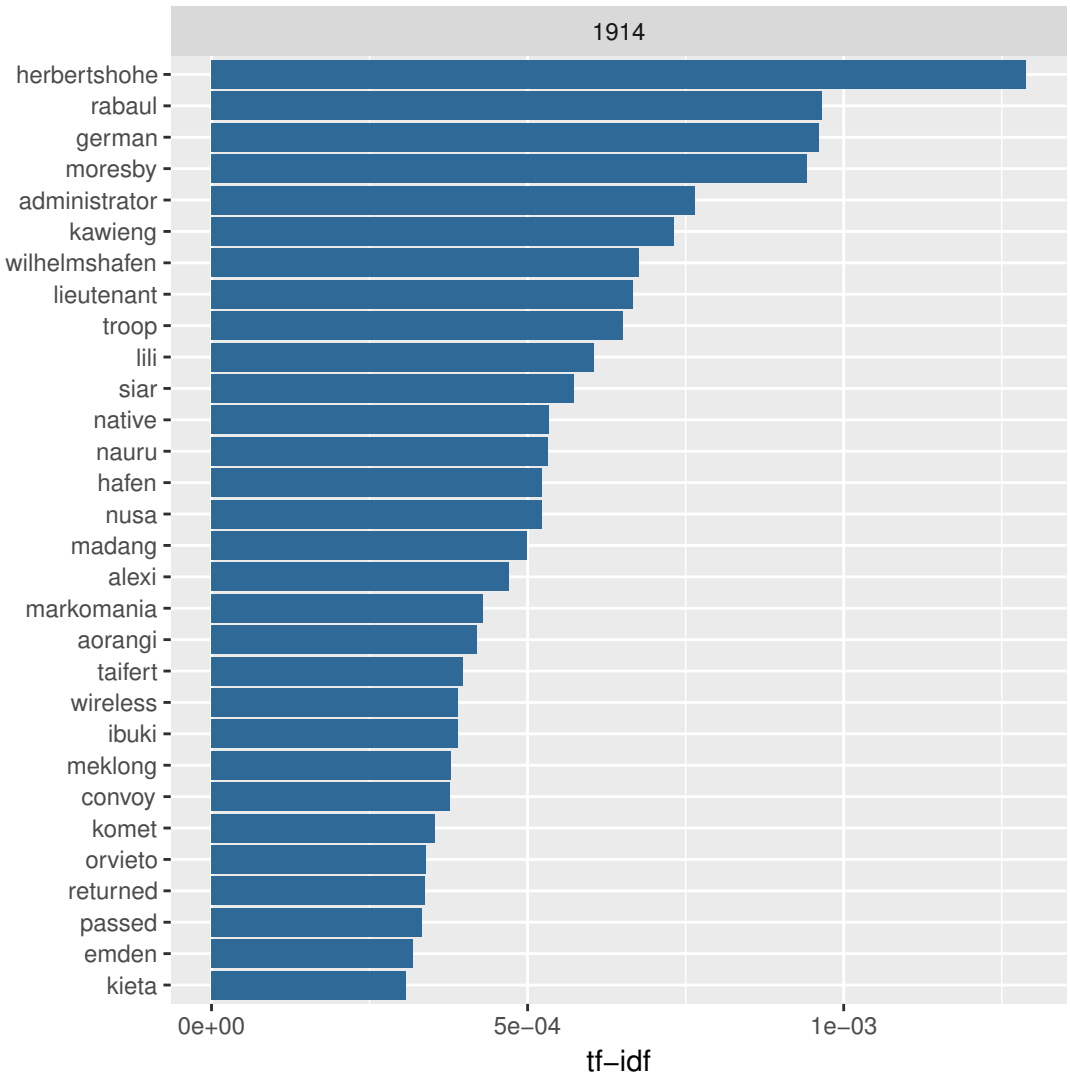


Figure 5.9: Words with highest 30 tf-idf scores for 1914.

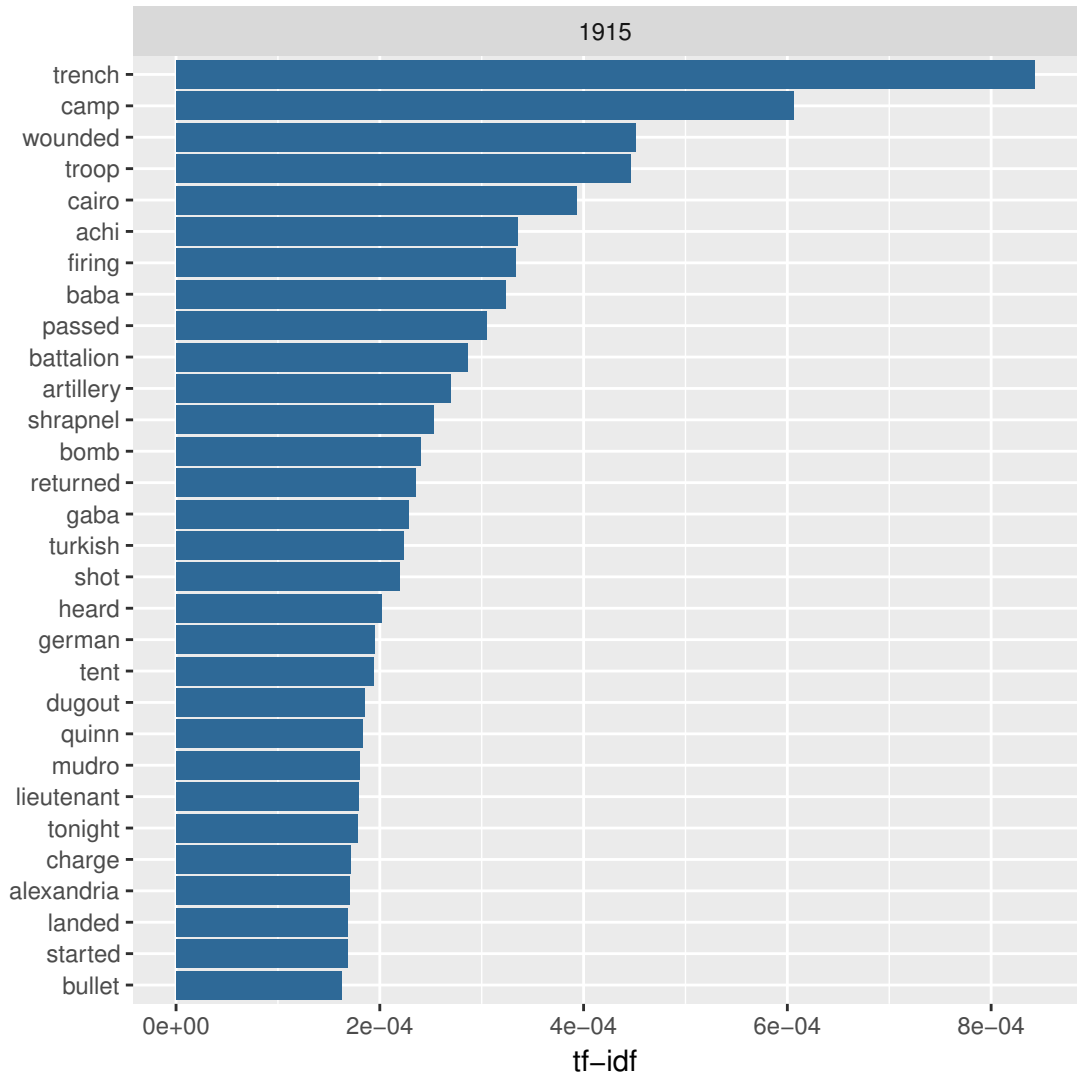


Figure 5.10: Words with highest 30 tf-idf scores for 1915.

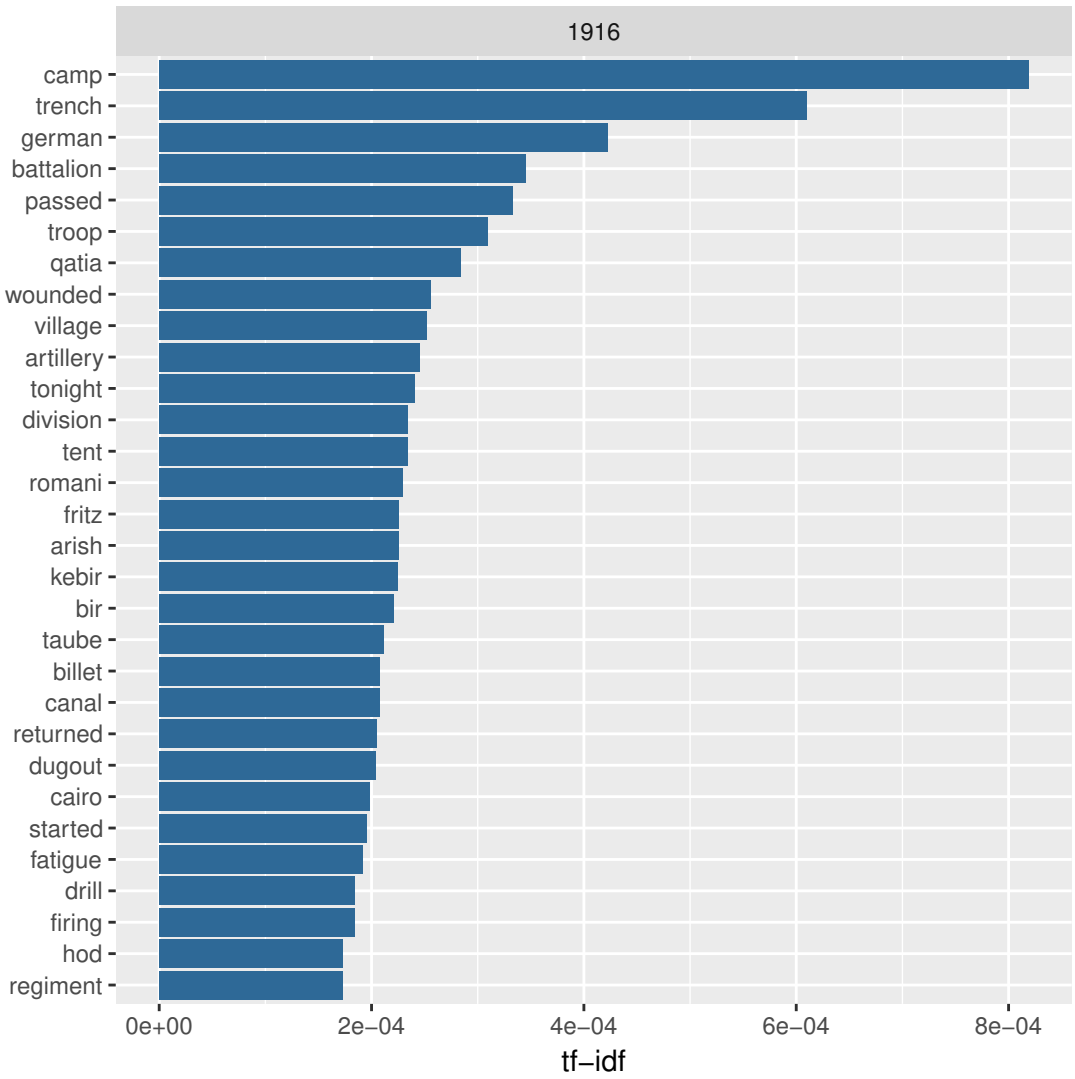


Figure 5.11: Words with highest 30 tf-idf scores for 1916.

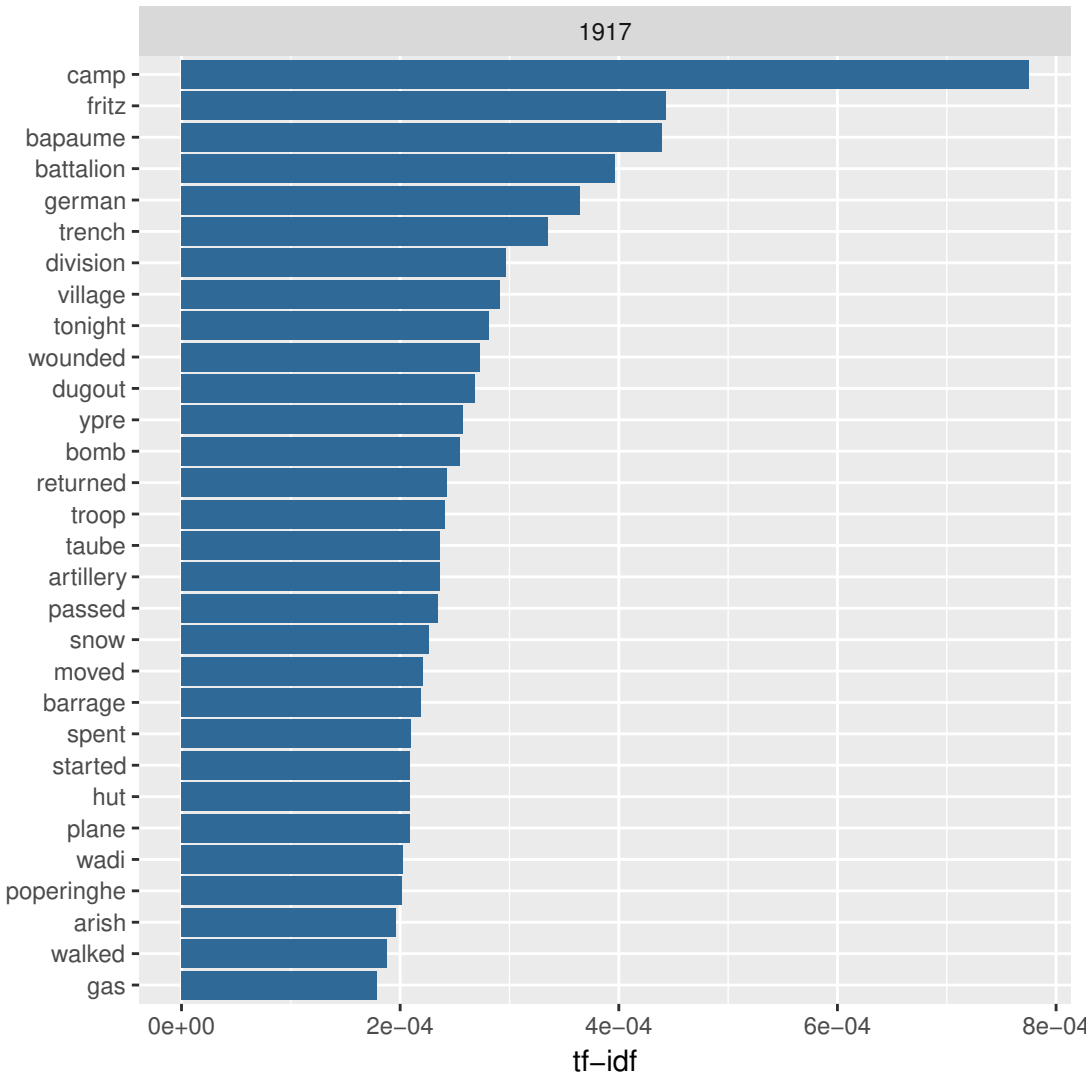


Figure 5.12: Words with highest 30 tf-idf scores for 1917.

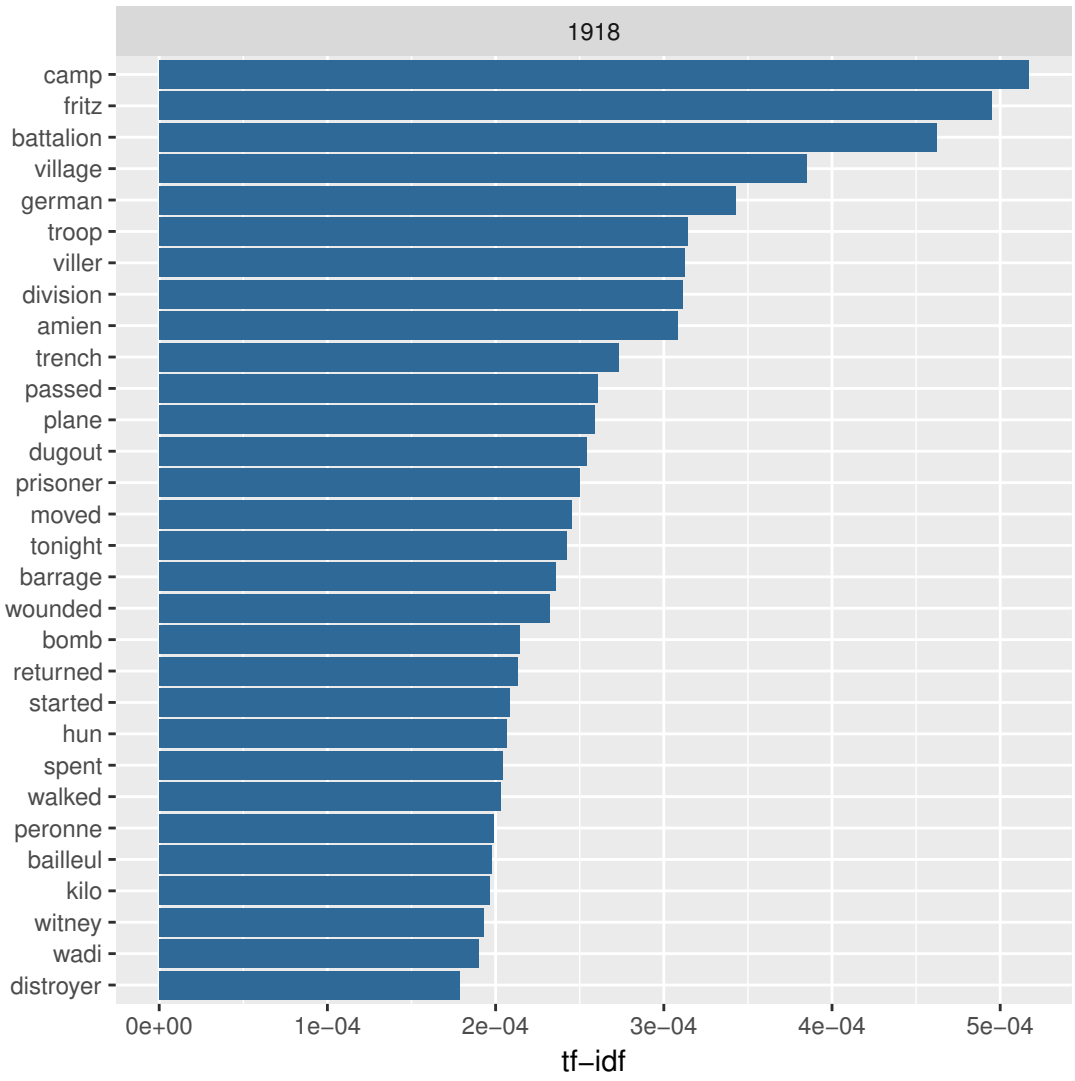


Figure 5.13: Words with highest 30 tf-idf scores for 1918.

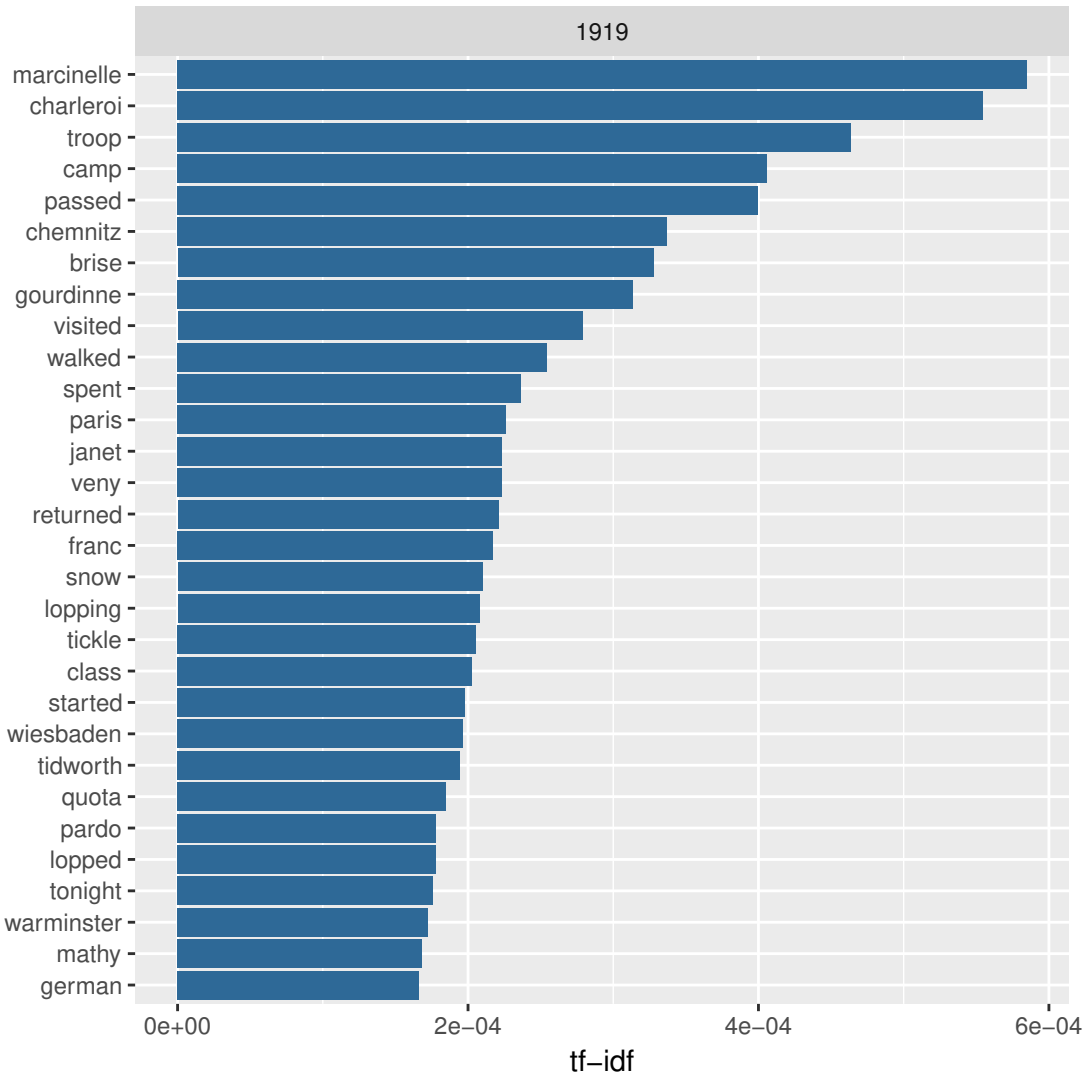


Figure 5.14: Words with highest 30 tf-idf scores for 1919.

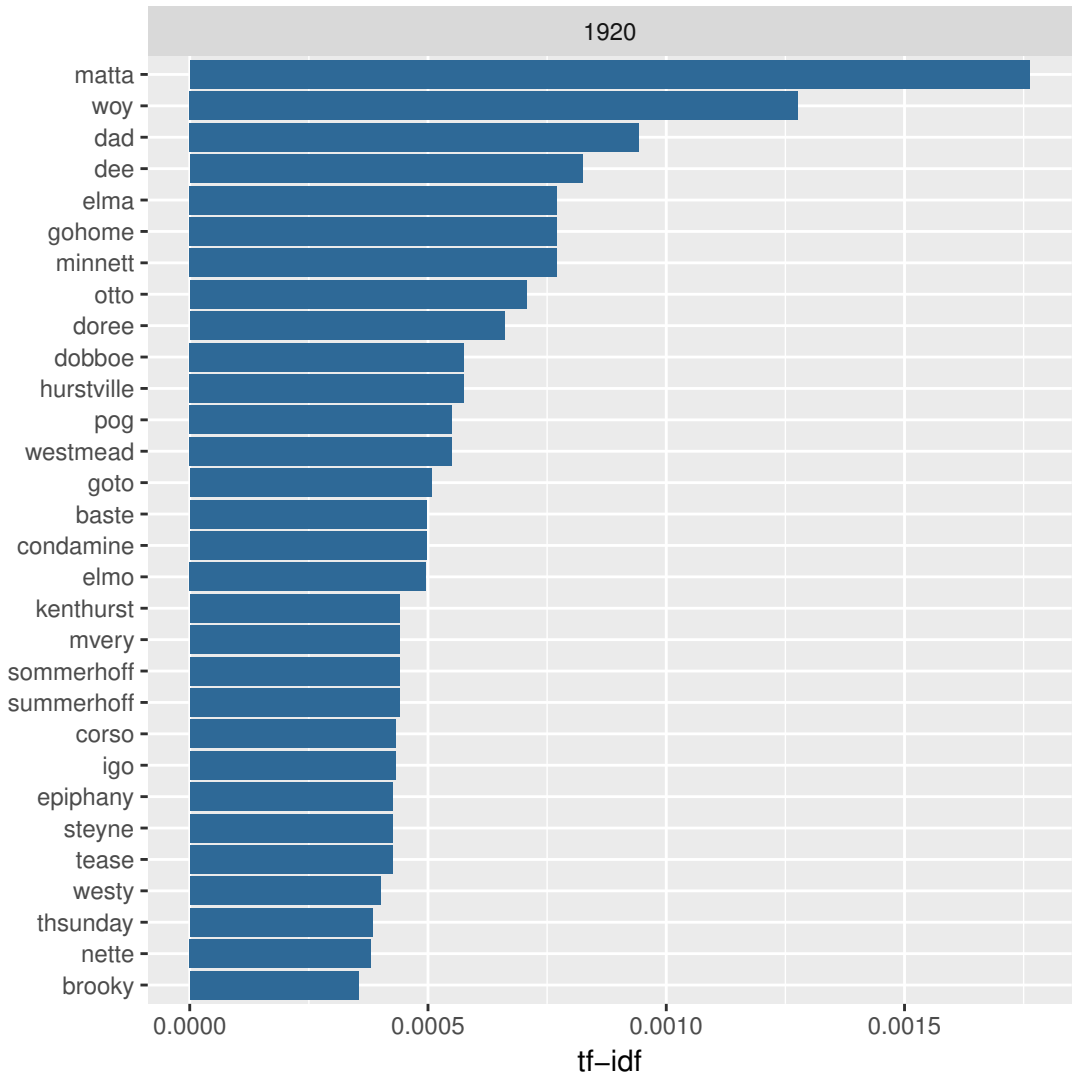


Figure 5.15: Words with highest 30 tf-idf scores for 1920 - 1923.

5.3 Topic Modelling

Topic modelling is a technique which was introduced to improve information retrieval methods [44]. However, it can be used in other applications including using it to understand themes throughout text documents [45]. Throughout this section we will discuss the theory behind several topic modelling techniques. These techniques will then be applied to our diaries in order to understand how the topics soldiers wrote about changed as the war progressed.

There are several different methods which have been used in topic modelling. However, they all make a *bag-of-words* assumption [17]. That is, they do not consider the sequential order of terms in a document, only their frequency. Further, they also assume that the order of the documents within the corpus is unimportant.

Currently, the primary method for topic modelling is LDA (Latent Dirichlet Allocation). However, this is derived from previous models of LSA and pLSA (Latent Semantic Analysis and probabilistic Latent Semantic Analysis). Alternatively, Gerlach et al. have suggested a network approach to topic models using SBMs (Stochastic Block Models). The theory behind the basic models of LSA, pLSA, LDA and SBMs will be discussed in Sections 5.3.1, 5.3.2, 5.3.3 and 5.3.4, respectively. In Section 5.3.5 we then perform our analysis using LDA.

Throughout this section the following notation will be used:

- Collection (corpus) of text documents, $\mathcal{D} = \{d_1, \ldots, d_N\}$,
- Vocabulary of all words, $\mathcal{W} = \{w_1, \ldots, w_M\}$.

5.3.1 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA), also known as Latent Semantic Indexing (LSI), was introduced by Deerwester et al. as an improvement over term-matching in information retrieval [44]. The theory behind LSA, which is summarised in this section, is based on Deerwester et al.'s paper [44] and Hoffman's paper [46].

There are two main problems with retrieving information by simply term-matching: synonymy and polysemy [44]. Synonymy refers to how people use many different terms to refer to the same thing, for instance, "happy", "joyful", "merry", etc., all describe positive emotions. Polysemy refers to the fact that a word may have several meanings, such as "keyboard", which could refer to the musical instrument or a computer keyboard. These problems mean that by using simple term-matching we will retrieve many documents which are un-related to what we are searching for. To overcome this, Deerwester et al. determined a method of finding relationships between terms and documents.

This is done by considering a term/document co-occurrence matrix, then analysing this using singular value decomposition (SVD) [44, 46]. Suppose we take our collection of text documents $\mathcal{D}$, and our vocabulary $\mathcal{W}$, then, by making a bag-of-words assumption, we can create a $M \times N$ co-occurrence matrix of counts $\mathbf{N} = (n(w_i, d_j))_{ij}$. Here, $n(w, d) \in \mathbb{N}$ gives the frequency of term w in document d. SVD is now used to decompose $\mathbf{N}$, into the product of three other matrices: $\boldsymbol{N} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices ($\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$) of the *left* and *right singular vectors* and $\boldsymbol{\Sigma}$ is a diagonal matrix of *singular values*. Note that SVD is "unique up to certain row, column and sign permutations" [44]. Hence, $\boldsymbol{\Sigma}$ is constructed such that the diagonal elements are positive and ordered by decreasing magnitude. These vectors are currently high-dimensional and sparse, however, we can convert these to low-dimensional, typically, non-sparse latent vectors by approximating the SVD. This is done by keeping the largest k singular values, and setting the rest to zero. The rows and columns in $\boldsymbol{U}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ which correspond to these zero singular values can then be deleted to form an approximation of $\boldsymbol{N}$: $\boldsymbol{N} \approx \hat{\boldsymbol{N}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^T$.

This approximate SVD can now be used to compare two terms, two documents and a word with a document [44]. To compare two terms we take the dot product between the corresponding two row vectors of $\hat{\boldsymbol{N}}$. Further, to consider the comparison of every pair of terms we calculate $\hat{\boldsymbol{N}}\hat{\boldsymbol{N}}^T$ which is a square symmetric matrix, and due to the properties of matrices in an SVD, we know $\hat{\boldsymbol{N}}\hat{\boldsymbol{N}}^T = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}^2\hat{\boldsymbol{U}}^T$. Similarly, two documents can be compared by taking the dot product of the corresponding column vectors of $\hat{\boldsymbol{N}}$. Once again, we can consider the comparison of every pair of documents by calculating $\hat{\boldsymbol{N}}^T\hat{\boldsymbol{N}} = \hat{\boldsymbol{V}}\hat{\boldsymbol{\Sigma}}^2\hat{\boldsymbol{V}}^T$. A term and a document can be

compared by simply considering the element in the corresponding row and column of $\hat{\boldsymbol{N}}$.

5.3.2 Probabilistic Latent Semantic Analysis (pLSA)

Probabilistic Latent Semantic Analysis (pLSA), also known as probabilistic Latent Semantic Indexing (pLSI) or the Aspect Model, was introduced by Hofmann [46] as an alternative to LSA. The theory behing pLSA, which is summarised in this section, is primarily based on Hoffman's paper [46].

pLSA introduces a latent variable (a variable which is not observed) $z \in \mathcal{Z} = \{z_1, \ldots, z_K\}$, which can be considered as "topics". Hence, we can define a joint probability model over $\mathcal{D} \times \mathcal{W}$ using the following:

$$\begin{aligned} P(d, w) &= P(d)P(w|d) \\ P(w|d) &= \sum_{z \in \mathcal{Z}} P(w|z)P(z|d). \end{aligned}$$

We can consider a document to be a "mixture of topics, where topics are a probability distribution over words" [16]. Based on this, the simple probabilistic method to generate a document can be described by the following generative model:

1. Choose the number of words in the document
2. Choose a distribution over topics
3. For each word in the document:
   (a) Randomly choose a topic from the distribution
   (b) Randomly choose a word from the probability distribution of words in that topic

Hence, to understand our documents we wish to know the probability of each topic in the documents ($P(z \mid d, w)$), and the probability distribution of words in each topic ($P(w \mid z)$). In order to get find these probabilities the Expectation Maximisation (EM) algorithm is used.

The EM algorithm can be used to maximise likelihood functions for problems where the likelihood is difficult to determine [47]. This algorithm assumes there is latent data which if included in the model would make the likelihood easier to calculate. The algorithm begins by initially guessing parameters before iterating over an expectation step and maximisation step to determine better parameters. This iteration continues until the parameters converge.

For pLSA the E-step equation is

$$P(z \mid d, w) = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum_{z' \in \mathcal{Z}} P(z')P(d \mid z')P(w \mid z')},$$

and the M-step equations are

$$\begin{aligned} P(w \mid z) &\propto \sum_{d \in \mathcal{D}} n(d, w)P(z \mid d, w), \\ P(d \mid z) &\propto \sum_{w \in \mathcal{W}} n(d, w)P(z \mid d, w), \\ P(z) &\propto \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d, w)P(z \mid d, w). \end{aligned}$$

This method allows us to understand documents in terms of the topics they are written about, however, it has some problems [17]. First, as the corpus of documents gets larger the number of parameters to estimate grows which can lead to overfitting. Overfitting is when a model fits the training data too well, causing it to be inaccurate when using that model on new data [48]. Also, in pLSA we are unable to assign probabilities to documents which are not in the training set [17].

5.3.3 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) was introduced by Blei et al. [17] as an extension to pLSA to address the problems of overfitting and missing data in the model. This is done by introducing a Dirichlet prior on the topic distribution [16, 17]. Hence, the generative model is now:

1. Choose the number of words in the document
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each word in the document:
   (a) Choose a topic $z \sim \text{Multinomial}(\theta)$
   (b) Choose a word w from $p(w \mid z, \beta)$

There are several assumptions made in this basic LDA model in order to simplify it [17]. First, the number of topics, k, and hence the dimension of the Dirichlet distribution, is assumed to be known and fixed. It is also assumed that the probabilities of words within topics are parameterised by a $k \times M$ matrix β where $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$ [17]. Also, a Dirichlet prior was chosen for simplicity as it "is in the exponential family ... and is conjugate to the multinomial distribution" [17]. Based on this generative model, the joint distribution is defined by [17]

$$p(\theta, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta) \quad = \quad p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) P(w_n \mid z_n, \beta).$$

There are now several methods which can be used to estimate the parameters in this model with the most commonly used method for LDA being Gibbs Sampling. An explanation of the Gibbs Sampler can be found in Casella and George's paper, *Explaining the Gibbs Sampler* [49].

There is no natural or commonly accepted way to determine the appropriate number of topics, k, in LDA. Choosing too many topics gives a model that is too complex and difficult to interpret [50]. By contrast, not having enough topics can make the model too coarse and hence make it too hard to accurately classify documents. Different methods to determine the number of topics have been suggested and we will focus on four of these.

Arun et al. [51] considers LDA as a matrix factorisation method which breaks a document-word frequency matrix into a topic-word matrix and a document-topic matrix. Their proposed measure is to determine the number of topics for which the Kullback-Lieber divergence between these matrices is minimised.

Cao et al. [52] considers topic density in order to determine the optimal number of topics, where the density of topic is based on the number of topics whose average cosine distance is within a chosen radius. This is based on the idea that the similarity within clusters should be large, whilst the similarity between clusters should be small, and that topics are equivalent to semantic clusters.

Deveaud et al. [53] determine the optimal number of topics by maximising the Jensen-Shannon divergence between pairs of topics.

Griffiths and Steyvers [54] determine the optimal number of topics by computing estimates of $P(\mathcal{W} \mid k)$ for varying k, where $\mathcal{W}$ are the words in the corpus and k is the number of topics. As $P(\mathcal{W} \mid k)$ is the likelihood in determining $P(k \mid \mathcal{W})$, we wish to find the number of topic which maximises this value.

Through the introduction of a Dirichlet prior LDA addresses both problems with pLSA. However, LDA still has some shortcomings [55]. First, whilst some methods to determine the number of topics have been suggested, there is no natural way to determine this. Also, whilst the introduction of a Dirichlet prior reduces the number of parameters there is still a large number of parameters to estimate which can cause overfitting. Finally, the only justification for introducing a Dirichlet prior is convenience and simplicity of the algorithm. However, this prior does not necessarily match properties of text.

5.3.4 Stochastic Block Models (SBMs)

To address the shortcomings of LDA, Gerlach et al. [55] proposed using hierarchical Stochastic Block Models (SBMs). This originates from a similarity between SBMs used for community detection and pLSA used for topic modelling. Community detection aims to identify groups of nodes within a network with similar connectivity patterns. Similarly to topic modelling, there are many different approaches to community detection, including modularity maximisation, SBMs and hierarchical SBMs, each of which is an improvement on the previous. Gerlach et al. show that SBMs are equivalent to pLSA for topic modelling, and hence hierarchical SBMs can be used for topic modelling.

5.3.5 Analysis

We wish to know how the topics soldiers wrote about changed over time. To do this we will use LDA and consider each document to be all the entries written in a particular month. First, we determine the optimal number of topics based on the methods described in Section 5.3.3. These were implemented using the `ldatuning` package [56] in R. The results from each method are given in Figure 5.16. Based on this, we find that Deveaud2014 gives the best results for 8 - 12 topics, Arun2010 gives best results for 6, 7 or 10 topics, Griffiths2004 gives best results for 8 or more topics, and CaoJuan2009 gives best results for 17 - 20 topics. Based on these results, we will perform topic modelling with 10 topics as this falls in the range of best parameters for three of the methods.

LDA can now be applied to our corpus using the `topicmodels` package [57] in R, using Gibbs Sampling with 10 topics and a seed of 1915. The 99 most probable words from each topic are given in Appendix F and based on these words we have given each topic a name which describes it. Hence, our topics are: *Everyday Life*, *War at Sea*, *Egypt*, *Gallipoli*, *In the Trenches (Beginning)*, *In the Trenches (Middle)*, *In the Trenches (End)*, *White Christmas*, *After the Armistice*, and *Home Again*.

We can now consider the proportion of each topic in each month over the course of the war. We have applied a rolling mean with $k = 5$ to these proportions before graphing them, and the resultant graph can be seen in Figure 5.17. For a series of data $y_1, \ldots, y_n$, then the rolling mean for each point, m_i, is found by taking the average of that point along with the l points either side, i.e., [58]

$$m_i = \frac{1}{2l+1} \sum_{j=-l}^{l} y_{i+j}, \quad i = l+1, l+2, \ldots, n-l.$$

Rolling means can be applied in R using the `rollmean` function in the `zoo` package [59]. In this function it is necessary to supply the width of the rolling window which is determined by $k = 2l + 1$.

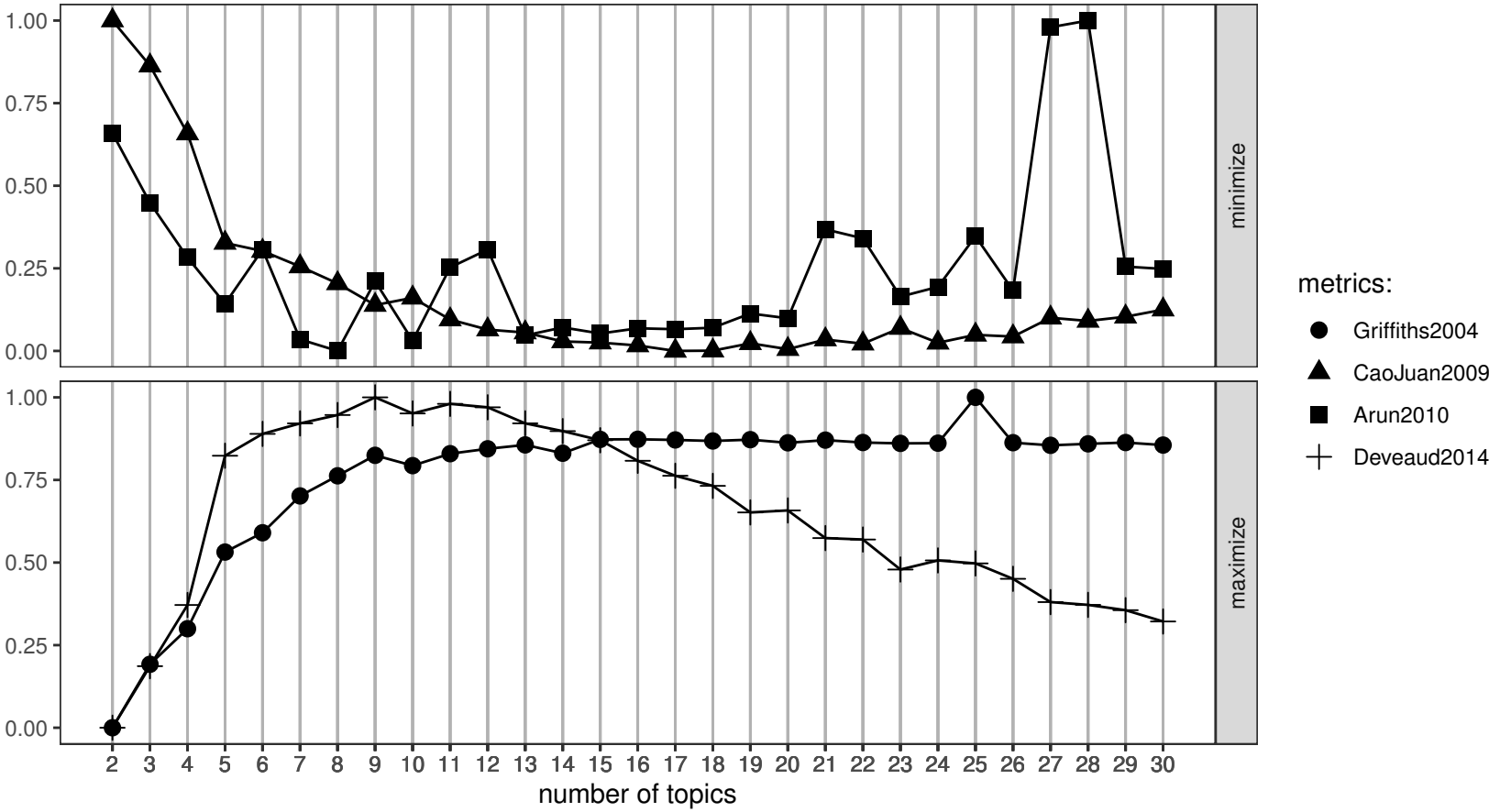


Figure 5.16: Results found by applying the four methods for determining the number of topics that were discussed in Section 5.3.3.

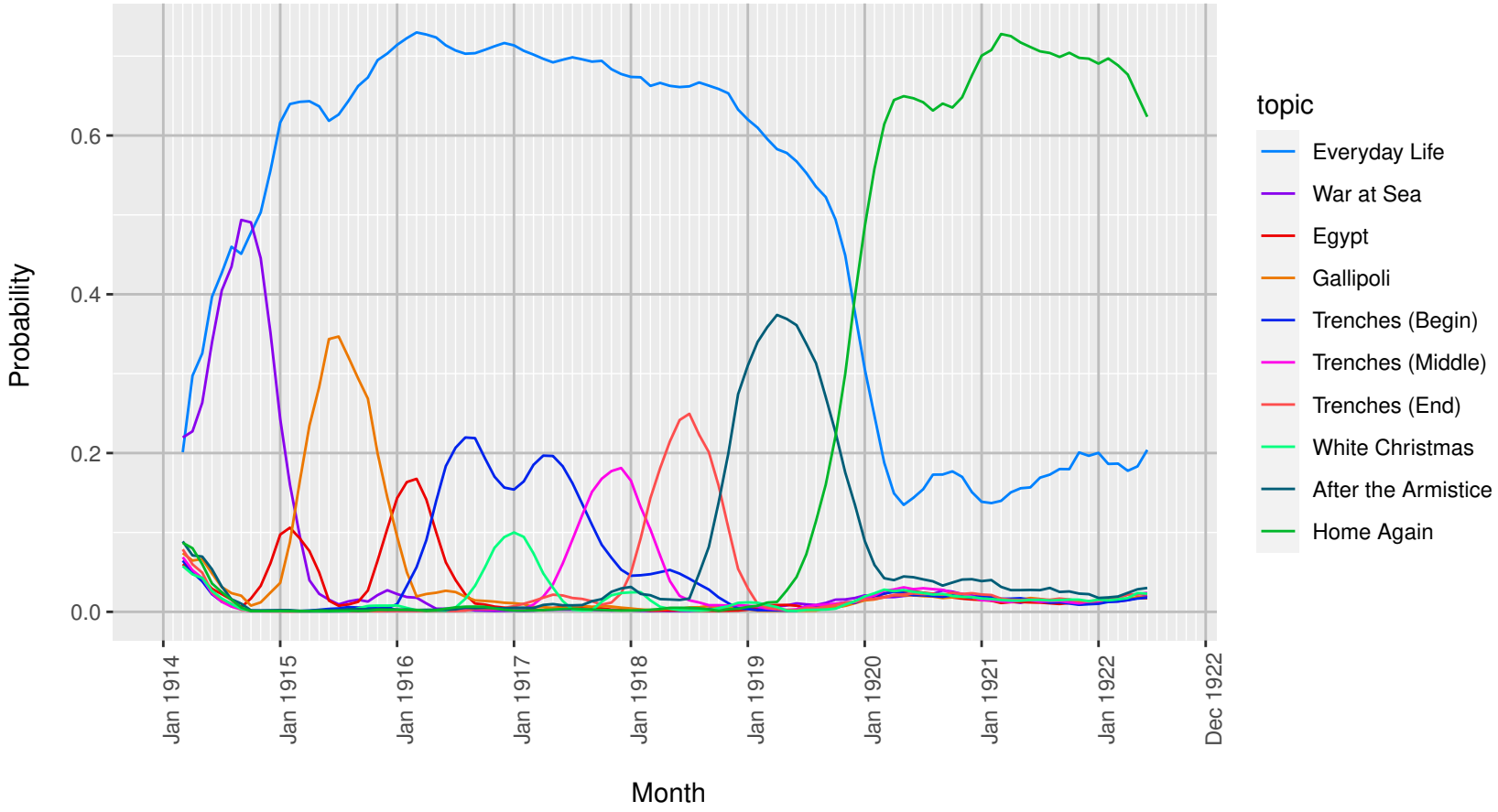


Figure 5.17: The proportion of each topic obtained from our LDA model, over time. Note that a rolling mean with $k = 5$ has been applied to each point.

From Figure 5.17 we first note that two topics, *Everyday Life* and *Home Again*, are the primary topics discussed throughout the diaries. Until the end of 1919 the *Everyday Life* topic is prominent, showing that whilst diarists did write about battles, training, travelling, etc., many of the entries focus on day-to-day activities. From 1920 onwards the *Home Again* topic is the most prominent, although *Everyday Life* is still present. This is to be expected as by this time the soldiers had returned home to Australia.

We will now focus on each of the other topics in turn discussing what time period they primarily spanned, their peaks and troughs, and what events they correlate with.

The *War at Sea* topic primarily spans from April 1914 to May 1916. This topic has a main peak around September 1914, corresponding to when Australia captured German New Guinea. This topic then dips in July 1915 before reaching a very small peak in December 1915. This second peak corresponds to the withdrawal of Australian forces from Gallipoli. This topic has probability almost zero elsewhere, as once troops were in Europe and the Middle East they used other modes of transport such as trains.

The *Egypt* topic spans from September 1914 to September 1916. This topic has two peaks in February 1915 and March 1916, with a trough in July 1915. The peaks correspond to periods when Australian soldiers trained in Egypt, with the dip in this topic being when Australian forces were sent to fight in the Gallipoli campaign.

The *Gallipoli* topic spans from October 1914 to March 1916, with a peak in June/July 1915. This roughly matches the span of the Gallipoli campaign, with the landing on 25th April 1915, withdrawal on the 20th December 1915, and many major battles, including Lone Pine, occurring in August.

The three *In the Trenches* topics altogether span from January 1916 to January 1919, covering the primary period that Australians were engaged in battles in Europe and the Middle East. Each of these topics have peaks, occurring in: August/September 1916 and April/May 1917 for the *Beginning* topic, December 1917 for the *Middle* topic, and June/July 1918 for the *End* topic. These peaks most likely correspond to specific battles, for example: the Battle of Romani (August 1916), the Second

Battle of Arras (April - May 1917), the Battle of Jerusalem (November - December 1917), and the Battle of Hamel (July 1918), as more troops were in the trenches fighting at those times.

The *White Christmas* topic has four peaks which span the European winters over December 1915, 1916, 1917 and 1918.

The *After the Armistice* topic spans from July 1918 to March 1920 with a peak in April 1919. After the armistice was declared in November 1918 many Australian soldiers had to wait in Europe for up to a year before being repatriated to Australia [60, 61].

5.4 Comments and Future Work

We have been able to use each of our three different techniques: word frequencies, tf-idf and topic modelling to gain an understanding about what soldiers wrote about during the war. Even though our results are only based on the data, through each technique we have overall seen events know from history. Interestingly, each of these techniques has shown different aspects of the war. Considering word frequencies of the entire text did not give specifics about any battles or places, only general terms to do with warfare. We were then able to go further and look at specific n-grams for subjects of interest, however, this restricts us to looking only at chosen n-grams. In tf-idf we see specific words that are unique to each year, and hence we see many locations and ship names, specific to events that happened in that year. Conversely, through topic modelling we see that most of the time they are actually talking about everyday things. We do still see battles, locations and ships etc., through the other topics. When comparing topic modelling and tf-idf we see that some of the results are the same, for instance, the *War at Sea* topic is very similar to the tf-idf for 1914, and likewise the *After the Armistice* topic matches the tf-idf for 1919. However, the tf-idf results for 1916 do not match the most prominent topic for that year: *In the Trenches (Beginning)*. The tf-idf results for 1916 show many words relating to locations and battles in the Middle East whilst the *In the Trenches (Beginning)* topic shows words relating to locations in Europe.

Using LDA breaks the diaries up into the various different stages of the war, however, does not seem to distinguish much between the three *In the Trenches* topics, or the battles in the Middle East and Europe. In the future we could consider using SBMs to try to improve results as for this method we do not need to know number of topics beforehand.

We will now use sentiment analysis to determine the soldiers emotions throughout the war.

Chapter 6: Sentiment Analysis

Sentiment Analysis, also known as opinion mining, attempts to understand the attitude or emotion of the author towards the subject of a text. Generally, sentiment analysis methods fall into three categories: dictionary based methods (DBMs), supervised learning methods and unsupervised (or deep) learning methods [62], and these will be discussed in Sections 6.1, 6.2 and 6.3, respectively. Our analysis focuses on using dictionary based methods and our results are given in Section 6.4. In Section 6.5 we discuss possible future work.

6.1 Dictionary Based Methods

Our analysis focuses on dictionary based methods (DBMs), which use a dictionary of terms with associated sentiment (valence) scores, known as a *sentiment dictionary* or *sentiment lexicon*, to determine the average sentiment of a document [62].

The average sentiment of text T using dictionary D is given by

$$s_D^T = \frac{\sum_{w \in D} s_D(w) f^T(w)}{\sum_{w \in D} f^T(w)},$$

where $f^T(w)$ is the frequency of word w in text T and $s_D(w)$ is the sentiment (valence) score of word w in dictionary D [62].

The main advantage of using DBMs is that they can be applied to text where there is no previous known information about its sentiment [62]. Also, by using DBMs one has the ability to know what words are contributing to the sentiment score. However, Reagan et al. [62] also note that DBMs have several disadvantages. First, DBMs must be applied to text much larger than a single sentence. Also, supervised learning methods trained on a sufficiently large training set for a corpus will always

perform better than DBMs applied to that corpus. Lastly, dictionaries cannot cover all words and their meanings. This means that common emotive words from a particular text may not appear in the dictionary and these dictionaries may rate words based on a different meaning than what is used in the text. For instance, "miss" would have a negative connotation in dictionaries. However, this is not the case when the word is used as a title.

Several different sentiment dictionaries exist but for this research we will consider the following: AFINN, ANEW, Hului, Loughran-Mcdonald, NRC, SenticNet, SentiWordNet, and Syuzhet. These dictionaries were chosen as they are freely available, primarily through the `lexicon` package [63] in R. In Subsections 6.1.1 - 6.1.8 we give a brief summary of each dictionary including when (and in some cases how and why) it was created, how many words it includes, the rating scale, and where it is available. Note that when we specify how many words a dictionary includes, this count also includes phrases given in the dictionary such as "too much fun".

6.1.1 AFINN

The AFINN lexicon was created in 2011 by Finn Årup Nielsen specifically for microblog-like content [64]. As such, it was designed to include obscene words and Internet slang such as "WTF" and "LOL" [64]. The AFINN lexicon includes a list of 2,477 words with each word having a valence score in the range $(-5, 5)$ where -5 is very negative and +5 is very positive [62, 64]. The word list was created based on several different sources and the valence of each term was manually chosen by the author [64]. The AFINN lexicon is available as part of the `tidytext` [2] package in R.

6.1.2 ANEW

ANEW was developed in 1999 by Margret Bradly and Peter Lang of the NIMH Centre for Emotion and Attention [65]. The aim was to provide ratings for a set of English words in terms of valence, arousal and dominance [65]. Each word was rated on a 9 point system, from pleasant to unpleasant (valence), calm to excited (arousal), and in control to dominated (dominance) [65]. The ANEW lexicon includes 1,034 words with each word having a valence score in the range $(1, 9)$ where 1 is very negative and 9 is very positive [62, 65]. The words in

this lexicon were rated by introductory psychology students (gender-balanced) who were each given 100-150 words and asked to rate them based on their immediate reaction to the word [65]. A csv containing the ANEW lexicon was obtained from Andrew Reagan's GitHub folder: `https://github.com/andyreagan/labMT-simple/tree/master/labMTsimple/data/ANEW`.

6.1.3 Huliu

The Huliu lexicon was created in 2004 by Minqing Hu and Bing Liu based on their work with customer reviews [66, 67]. The Huliu lexicon includes 6,874 words with possible sentiment values in the set {1, 0, -1.05, -1, -2} [66]. A sentiment value of -2 means that the phrase is always negative, e.g., "too much fun" or "too much evil" [66]. The Huliu lexicon is available through the `lexicon` package [63] in R.

6.1.4 Loughran-McDonald

The Loughran-Mcdonald lexicon was initially created in 2011 by Tim Loughran and Bill McDonald based on words used in a sample of financial summaries from U.S. companies (10-K forms) from 1994 - 2008 [68]. This lexicon contains 2,702 words with a sentiment rating of either -1 (negative) or 1 (positive) [66]. The Loughran-McDonald lexicon is available through the `lexicon` package [63] in R.

6.1.5 NRC

The NRC lexicon was created in 2010 by Saif Mohammad and Peter Turney [66]. The words in this lexicon were chosen by comparing the words and phrases from the Maquarie Thesaurus with those that occur frequently in the Google n-gram corpus, the General Inquirer, and the WordNet Affect Lexicon [69]. These words were then rated using the crowd sourcing website: Amazon Mechanical Turk [69]. The NRC lexicon contains 5,468 words with with a sentiment rating of either -1 (negative) or 1 (positive) [66]. The NRC lexicon is available through the `lexicon` package [63] in R.

6.1.6 SenticNet

The SenticNet 4 lexicon was created in 2016 and contains 23,626 words with associated sentiment scores in the range $(-1, 1)$, with -1 being the most negative and +1

being the most positive [62, 66, 70]. The SenticNet lexicon is available through the `lexicon` package [63] in R.

6.1.7 SentiWordNet

The SentiWordNet 3.0 sentiment dictionary was developed in 2010 by automatically annotating WordNet synsets [71]. WordNet originated at Princeton University in 1986 and is a collection of synsets (synonym sets) [72]. Each synset contains words which express the same concept, but are not necessarily interchangeable [72]. In the SentiWordNet lexicon each synset used is given a positive (Pos), negative (Neg) and objective (Obj) score [71]. This lexicon is available through the `lexicon` package [63] in R. Note that the SentiWordNet lexicon available through this package only contains one sentiment score per word which was determined by taking the difference between its postive and negative score, i.e., Pos - Neg. This lexicon contains 20,093 words with sentiment scores in the range $(-1, 1)$ [62, 66].

6.1.8 Syuzhet

The Syuzhet lexicon was developed in 2017 by the Nebraska Literacy Lab under the direction of Matthew Jockers as part of the `syuzhet` package in R [63, 73]. This lexicon contains 10,738 words with sentiment ratings in the range $(-1, 1)$ and is available through the `lexicon` package [63] in R.

6.2 Supervised Learning Methods

Supervised learning methods use statistical techniques to determine the sentiment of a data set based on labelled training data [74, 75]. This training data is usually manually rated as either positive, negative or neutral. Techniques such as Naive Bayes, Maximum Entropy, Support Vector Machines or Stochastic Gradient Descent can then be applied to determine the sentiment of the remainder of the data set.

The disadvantage of supervised methods is the requirement to have a labelled training set, which is not always possible. Since the documents on which sentiment analysis is used can be highly varied it is necessary to have training data specific to the documents being analysed [62]. This can be very time consuming to obtain and hence is often not practical.

6.3 Unsupervised Learning Methods

Unsupervised learning methods require no labelled data set. Several methods to do this have been proposed, for instance, Lin and He [76] proposed a method called joint sentiment/topic model (JST), based on LDA, to determine sentiment and topics simultaneously. This is done by adding an extra step into the generative model for LDA. In JST we first randomly choose a sentiment label, and then choose a topic distribution, and subsequently a word, from the sentiment distribution.

6.4 Analysis

To determine the sentiment of our diaries over time we used DBMs, over the data for each month. DBMs were chosen as we had no labelled data and also as they are simpler to implement than unsupervised methods.

Before determining the sentiment of our diaries we first consider what percentage of unique words in our diaries are covered by the sentiment dictionaries. We will also compare to this to the percentage of unique words from the Brown Corpus that appear in the sentiment dictionaries. The Brown Corpus contains a collection of documents that were printed in the United States in 1961 [77]. The collection contains 1,006,770 words, including 45,215 unique words. This is less than our diary collection which contains 9,266,353 words, including 84,955 unique words. The frequencies of words in the Brown Corpus was obtained using the `zipfR` package [78] in R. Table 6.1 gives the percentage of unique words from our World War I diaries and the Brown Corpus that appear in each of the sentiment dictionaries. From this table we note that approximately twice as many unique words from the Brown Corpus are covered by the sentiment dictionaries then the unique words from our diary corpus, even though there are less unique words in the Brown corpus. This suggests that the dictionaries may not be the best source for determining the

sentiment of our diaries. However, due to their simplicity we will use them for this research.

	Percentage (%)	
Dictionary	**WW1 Diary Corpus**	**Brown Corpus**
AFINN	2.03	4.35
ANEW	1.14	2.12
Huliu	5.09	9.86
Loughran-Mcdonald	1.59	3.80
NRC	5.49	10.09
SenticNet	14.71	26.67
SentiWordNet	7.83	14.01
Syuzhet	8.43	16.85

Table 6.1: The percentage of unique words from our World War I diaries and the Brown Corpus that appear in each of the sentiment dictionaries used.

In order to compare the results of different dictionaries we need them all to be on the same rating scale. We chose to convert them all to the scale -1 (very negative) to +1 (very positive), with 0 being neutral as five of the dictionaries already used this scale. The other dictionaries were converted to this scale using the formula:

$$x_{\text{new}} = \left(\frac{\max_{\text{new}} - \min_{\text{new}}}{\max_{\text{old}} - \min_{\text{old}}}\right)(x_{\text{old}} - \min_{\text{old}}) + \min_{\text{new}}, \tag{6.4.1}$$

where x_{old} and x_{new} are the old and new value, respectively, $[\min_{\text{old}}, \max_{\text{old}}]$ is the old value range, and $[\min_{\text{new}}, \max_{\text{new}}]$ is the new value range. An exception to this is the Huliu lexicon in which any word with a sentiment score of -2 was converted to a score of -1 (representing very negative).

After converting the rating scales of our dictionaries we were then able to graph the sentiment over time. Before graphing the sentiment we applied a rolling mean, with $k = 5$, to each dictionary. Figure 6.1 shows the sentiment over time of each of the dictionaries. Note that no data is shown after 1922 as only a few months had data, and hence a rolling mean could not be calculated across 1923.

From Figure 6.1 we can see that five of the dictionaries: AFINN, Huliu, Loughran-Mcdonald, NRC, and Syuzhet, follow the same general pattern. Based on Table 6.1, the other three dictionaries do not cover either more or less words in our corpus than the others, suggesting that these dictionaries may just contain a different selection of words. We also note there is a lot of variability in the sentiment after 1919 and in the first half of 1914. This is due to the small amount of data available for this time as seen in Figure 4.24. Hence, when the sentiment is calculated for these time periods it will only be based on a very small amount of words each month leading to large variability overall. Based on this, from now on we will only consider our sentiment between August 1914 and December 1919.

We obtain an overall pattern for our sentiment by taking the average of our eight sentiment dictionaries over this period. Figure 6.2 shows our average sentiment over time after a rolling mean, with $k = 5$, has been applied.

From Figure 6.2 we note that our average sentiment score between August 1914 and December 1919 is always positive (above zero). However, our average sentiment score never goes above 0.2, showing that whilst the diaries are positive, overall they never show a very positive sentiment. From the graph we also see there are several peaks occurring in: January/February 1915, February/March 1916, June/July 1917, January 1918, April/May 1919, and November 1919. There are also dips in sentiment occurring in June/July 1915, December 1916/January 1917, October 1917, July 1918, and September 1919.

To understand why these peaks and dips in sentiment occur we compare our average sentiment scores with our topic proportions over time. We exclude the *Everyday Life* and *Home Again* topics as they have consistently high proportions over large periods of time and hence cannot affect the sentiment. This comparison is graphed in Figure 6.3.

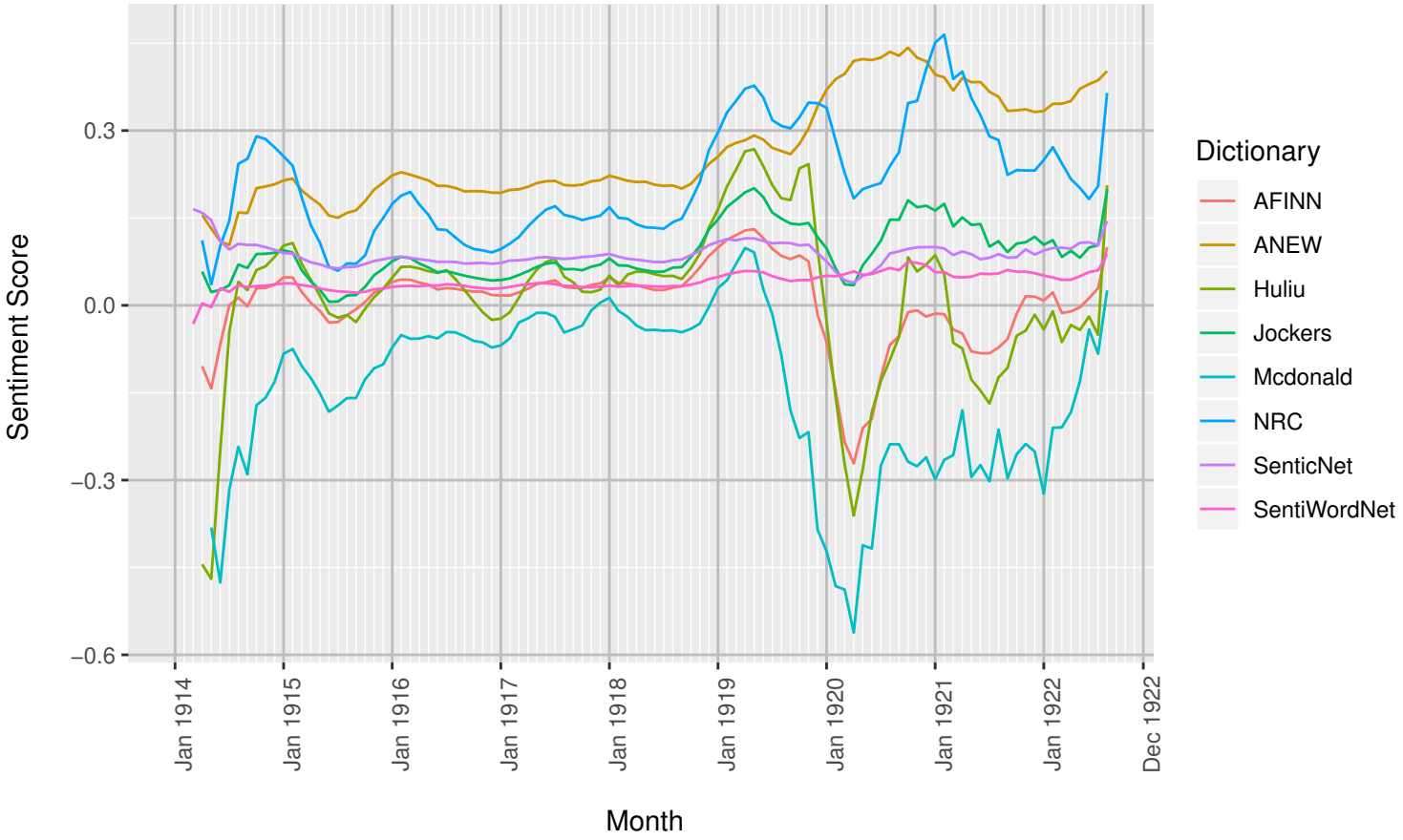


Figure 6.1: Sentiment scores over time for the eight dictionaries: AFINN, ANEW, Huliu, Loughran-Mcdonald, NRC, SenticNet, SentiWordNet, and Syuzhet. Note that before graphing we have applied a rolling mean, with $k = 5$, to each of the dictionaries.

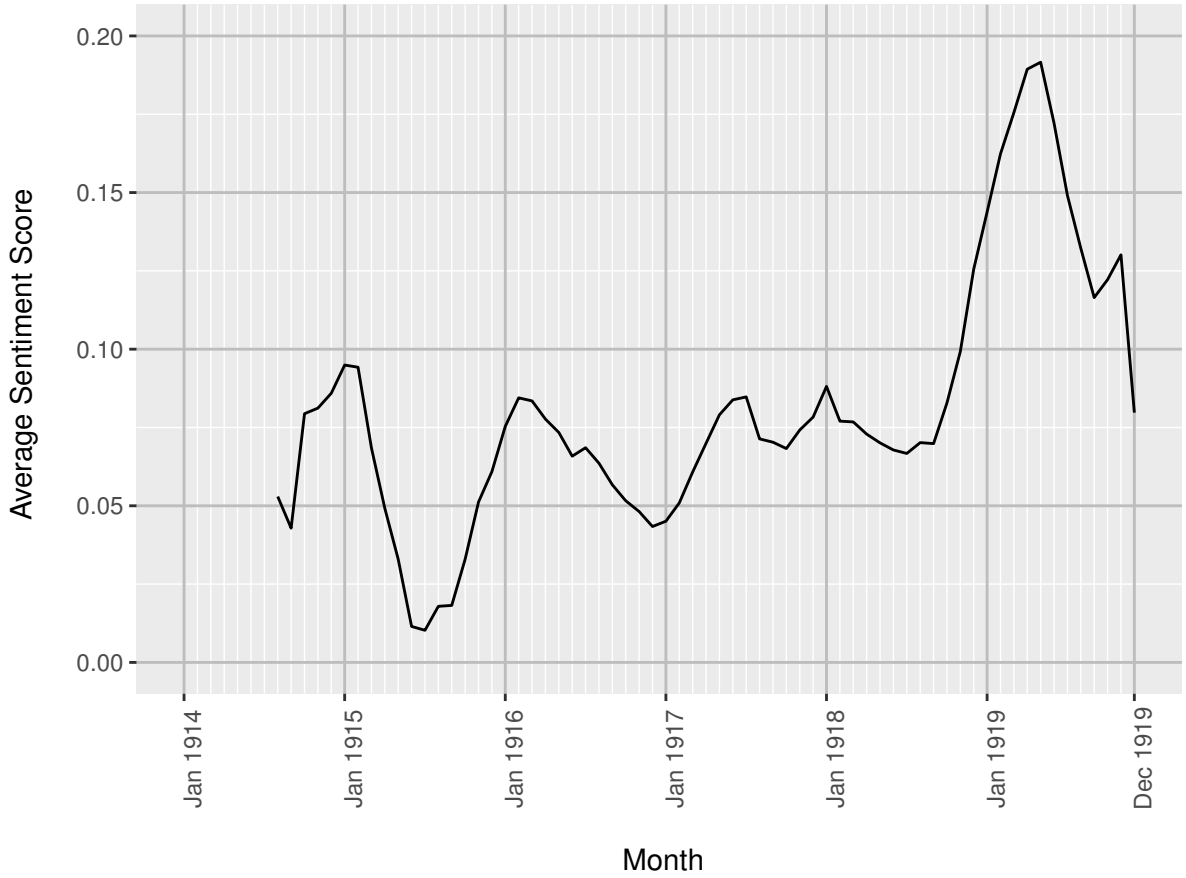


Figure 6.2: Average sentiment scores over time. Note, we do not show sentiment scores before August 1914 and after December 1919 due to a lack of data for these time periods leading to large variability. Before graphing we have applied a rolling mean, with $k = 5$.

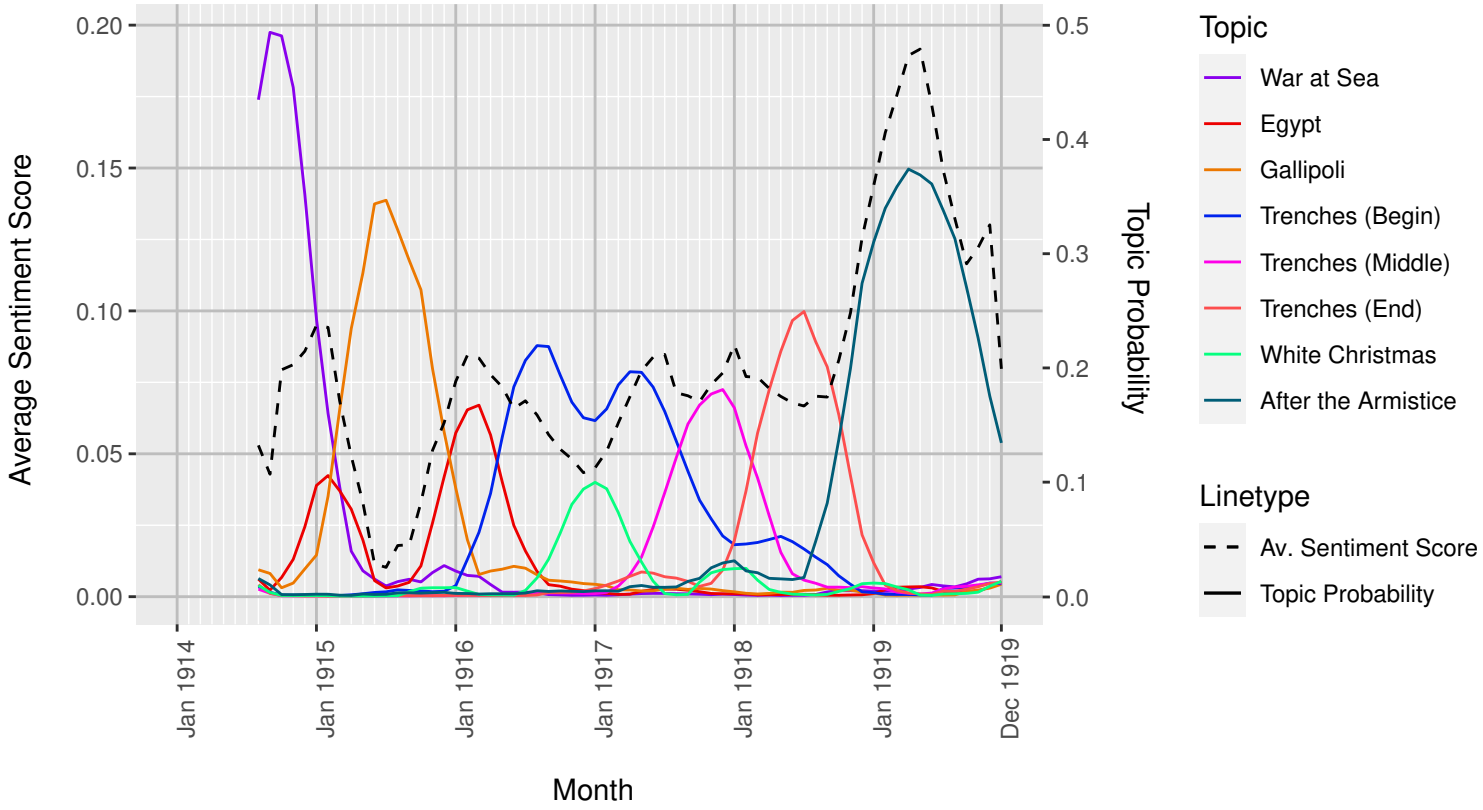


Figure 6.3: Average sentiment analysis scores compared to the topic probabilities (except the *Everyday Life* and *Home Again* topics).

In Figure 6.3 we see that some of the peaks and dips in our sentiment correspond to peaks in certain topics. In January/February 1915 and February/March 1916 we see peaks in our sentiment corresponding with the peaks in our *Egypt* topic. This is possibly due to the excitement of soldiers arriving in Egypt for training as they are in a new country and are excited about proving themselves in battle. Whilst in Egypt the men were allowed small trips into Cairo and around the pyramids, as seen in the following:

> "Climbed to the top of the Great Pyramid of Ghiza ... One got a very good view from the top of the Pyramid ... A sight which I dont think that anyone can see unless they go to Egypt to see it. The Sphinx simply fascinated one.", James McKenzie

In June/July 1915 we have a dip in sentiment corresponding to a peak in the Gallipoli topic. For many this would have been their first experience in battle, and seeing the true horrors of war. For instance, Thomas Munro writes:

> "It is an awful sight to see the dead and wounded on both sides, lying out and being walked on, no possibilite to bring them in or bury them, Some of our men have been out there a month and are still there. The stench would knock you down."

In December 1916/January 1917 there is a dip in sentiment corresponding to the largest peak in the *White Christmas* topic. In Appendix F we note that the word "miserable" is in the top 30 words of the *White Christmas* topic, and through close reading of the diaries we note that over these two months there are several comments regarding the cold and wet weather making things miserable. For example:

> "The weather is very miserable, dull gloomy skies with a few scattered showers during the day and at night it rains fairly constant ...", Archie Barwick

> "It is a very miserable existence now for one cannot light a fire after dark it gets very chilly in the morning & evening.", Maurice Evans

> "Raining heavily all day which made the conditions more miserable. The mud & slush is terrible.", Langford Colley-Priest

Hence, this dip in sentiment is primarily related to the cold and wet weather. Also, whilst some men had a good Christmas, others did not due to weather, lack of supplies, illness and being stuck on the front line. The contrast between these Christmases are seen in the following quotes:

> "Christmas dinner and tea were very merry, the rations being supplemented by a lot of luxuries ... also by plum pudding ... ", Hector McLean

> "A good many comforts for Christmas are arriving.", Thomas White

> "At mid-night a great crowd of us, sang Christmas Carols outside the Officers Quarters, a great joke half the crowd having had a drop too much.", Langford Colley-Priest

> "Christmas-day To-day we were given the best meal that I have had since I joined the A.I.F. At dinner we were given Pork & green peas, veal & plum Pudding.", Henry Parsons

> "Christmas week was very miserably spent on account of the rain wind & cold.", Gordon Macrae

> "After Christmas we was only allowed 1lb of bread per man owing to shortage of flour ... At this time typhus fever is very bad here especially amongst the Russians", Herbert Brown

> "Cold, miserable & hungry, we filed up to the cook house for our "Christmas dinner" of Bully beef Stew and buscuits [sic], as our rations were not yet to hand and our Christmas comforts were delayed somewhere.", Tom Taylor

> "Will go up and occupy the front line ... orders were gravely read, that there is to be no fraternisation with the enemy on Christmas Day ... Got some rum when I returned to the dugout, but our Rations have been lost so there was nothing to eat except a piece of dry bread and "Somme" mud. This, and the memories of other Christmas dinners I had eaten; was my Christmas dinner this year.", William Nicholson

Lastly, there is a peak in sentiment corresponding to peak in the *After the Armistice* topic. It is expected that after the armistice was signed that the sentiment of the soldiers would increase as the fighting was now over. This is further strengthened by the fact that it took up to a year for some to return home and whilst waiting for repatriation back to Australia soldiers spent their time travelling around France and Britain, attending sport matches and concerts as well as receiving vocational training from the AIF (Australian Imperial Force) [60, 61].

> "I then went to the Shakespeare Memorial Theatre and thoroughly enjoyed "As You Like It" by the Nigel Playfair Company, in quite a unique setting.", Thomas White
>
> "There was a Concert after tea in the Sydney Gardens, an Orchestral Band & some singing. I enjoyed it very much, out in the open air on a lovely Evening, in the Pretty Garden. ", Thomas Cleary

The other peaks and dips in sentiment are most likely related to specific battles or events which are not clearly seen through topic modelling. For instance, the dip in October 1916 may be related to attacks during the Third Battle of Ypres. In order to know exactly what these peaks and dips in sentiment are related to the diary entries for the relevant months would need to be manually read. However, due to the large amount of text involved this is out of the scope of this thesis.

6.5 Future Work

Our analysis has focused on using DBMs to understand the emotions of soldiers throughout the war. However, there are several things we can do in future. Each of our dictionaries were created since 1999, with some of them built for specific domains, such as micro-blog like content or financial documents. This does not take into account any language or use of words specific to wartime, or the possible language and word changes that may have happened in the 70 or more years between the war ending and these dictionaries being created. We need to investigate how language and word changes affect sentiment analysis results. The need to examine these dictionaries for their applicability to World War I text is further seen when we consider how many words in our corpus appear in the dictionaries. In Table 6.1 we saw that each of these dictionaries only cover a small percentage of words from our diaries corpus, especially when compared to a source such as the Brown Corpus,

suggesting that some of them may not be very applicable to World War I diaries. In this thesis we did not use supervised or unsupervised learning methods but these could be explored in future to see if they are more applicable then DBMs. Since all three methods have been developed more recently and primarily tested on product/movie reviews or social media, it would be worth determining how accurate they are on various historic documents.

Chapter 7: Conclusion

Throughout this research we have used mathematical and computational techniques to analyse Australian World War I diaries. Our aim was to understand what the soldiers wrote about and how they felt as the war progressed. More specifically, we aimed to create a technique to extract dates from diaries, as well as use both topic and sentiment analysis to analyse the text.

We began in Chapter 2 by giving a brief background of why Australia was involved in World War I and the impact of the war on Australia, both in terms of casualties and helping shape Australia's national identity. We also explained the advantages in using distant reading methods as opposed to more traditional close reading.

In Chapter 3 we discussed how the data was obtained and the three stages required to clean it. These stages were converting our raw data into a single text file per document with a metadata table giving identifying information for each document, extracting dates from each document, and removing numbers, punctuation and stop words as well as singularising words.

Chapter 4 then explained our date extraction process in more detail, including the use of regular expression to extract raw dates, and using an optimisation program to overcome problems found within our raw dates. These problems included missing information, mistaken information, and non-entry dates. After extracting dates we were able to convert our data into a date/entry format, allowing us to analyse our data over time.

Our first objective was to determine what the soldiers wrote about during the war. In Chapter 5 we considered three different methods to determine this. These were: word frequencies, tf-idf (term frequency - inverse document frequency) and topic

modelling.

By considering word frequencies we were able to see an overview of the collection as a whole. We categorised the top 100 most frequent words as either common, homonyms, or war related, and found that common words were the most frequent. Further, we were able to investigate particular subjects by considering the frequency of n-grams associated with them. Both tf-idf and topic modelling highlighted various stages of the war including: the use of ships in warfare, training in Egypt, the Gallipoli campaign, the fighting in Europe and the Middle East, and travelling and leisure activities after the armistice was signed. However, these techniques also had differences in what locations they showed as well as when the battles in the Middle East occurred. Tf-idf analysis shows the battles in 1916, whilst topic modelling shows those that occurred in 1917 and 1918.

Finally, in Chapter 6 we determined how the soldiers felt during the war using sentiment analysis. We explained the three different methods for determining sentiment: dictionary based methods (DBMs), supervised learning methods and unsupervised learning methods. We then used DBMs to determine an average sentiment for our diaries over the course of the war and compared this to our topic modelling results to determine why the sentiment peaked or dipped at certain times.

The use of distant reading techniques to find patterns within our corpus has allowed us to analyse a very large amount of text which would not be possible through traditional close reading. This is one of the concepts which form the basis of the emerging field of digital humanities in which mathematical and computational techniques are applied to traditionally humanities-based data. Through the use of regular expression and optimisation we were able to develop a new technique to extract dates from diaries, and correct any problems within these extracted dates. Through topic analysis we were able to determine what topics the soldiers wrote about over time. We were then able to compare this with our sentiment analysis results, giving us a quantitative view of the soldiers emotions throughout this period. This primarily confirmed what was already known about the experiences of Australian soldiers in World War I. It also allowed us to observe possible problems in this process when applied to historic data. This is a foundation, which in the future will allow us to examine these techniques for their accuracy on historic text.

There are several things that were brought up during this thesis that could be addressed in future work. These are listed below.

- Other libraries around Australia have been digitising their World War I diary collections. These could be added to the State Library of NSW's collection so that the experiences and thoughts of more people could be heard. However, it is important to note that these collections are not as large as the one held by the State Library of NSW and it is likely that the libraries have used different methods to transcribe the documents and create metadata, requiring a large amount of data cleaning.
- The metadata created for this corpus by extracting information from the titles of the documents could be improved by comparing it against the State Library of NSW's catalogue. This was not done in this work as it would take time and was not necessary as all the required metadata was extracted from the titles.
- When cleaning the data we could consider methods to stem words and correct spelling mistakes.
- When extracting raw dates we could consider other REs or methods such as named entity recognition (NER) to improve this process.
- We could consider adding the possibility of including a known end date into our optimisation program.
- We could test how the combination of problems affects our date extraction program by running simulations where two or more problems are combined.
- We could manually transcribe a subset of diaries so that we are able to fully test whether our extraction and optimisation methods are accurate.
- When performing topic modelling, we could consider using stochastic block models rather than LDA as SBMs do not require the number of topics to be known beforehand.
- We can look into using supervised and unsupervised methods to perform sentiment analysis and how accurate these three sentiment analysis methods are on World War I diaries, and more generally historic documents, given the changes in language that have occurred since these documents were written.

- Our analysis showed a slightly positive sentiment throughout the entire war, most likely as the majority of what the soldiers wrote about was everyday activities. We could perform aspect-based sentiment analysis to determine the sentiment for different types of events, e.g., battles, training and leave.
- We could compare the diaries to other sources from the time, i.e., newspapers or C. E. W. Bean's official history of Australia in World War I to see how the war was portrayed differently in these public sources compared to personal diaries.

The code used for this project and the simulation results from Chapter 4 can be found in the GitHub repository:

`https://github.com/AshleyDennisHenderson/Analysing-Australian-WW1-Diaries`

An interactive web app was created to allow users to explore the diaries using the techniques described throughout this thesis. This web app can be located using the following link:

`https://ashleydennis-henderson.shinyapps.io/Analysing-WW1-Diaries/`

Bibliography

[1] J. G. Royde-Smith and D. E. Showalter, "World War I," 2020. [Online]. Available: https://www.britannica.com/event/World-War-I

[2] J. Silge and D. Robinson, "tidytext: Text Mining and Analysis Using Tidy Data Principles in R," *JOSS*, vol. 1, no. 3, 2006. [Online]. Available: http://dx.doi.org/10.21105/joss.00037

[3] C. Craham, I. Milligan, and S. Weingart, *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College press, 2016.

[4] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann, "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges," in *Eurographics Conference on Visualization (EuroVis)*, R. Borgo, F. Ganovelli, and I. Viola, Eds., 2015.

[5] M. Caulfield, *The unknown Anzacs : the real stories of our national legend : told through the rediscovered diaries and letters of the Anzacs who were there*. Sydney: Hachette Australia, 2013.

[6] P. Cochrane, "'Diamonds of the Dustheap': Diaries from the First World War," *Humanities Australia: The Journal of the Australian Academy of the Humanities*, no. 6, pp. 22 – 33, 2015.

[7] ——, "Oh God what a fight," *The Australian*, 4 2011. [Online]. Available: https://www.theaustralian.com.au/arts/books/oh-god-what-a-fight/news-story/223f896c59318cdc1452c465d3767588

[8] ——, "The past is not sacred: the 'history wars' over Anzac," 4 2015. [Online]. Available: https://theconversation.com/the-past-is-not-sacred-the-history-wars-over-anzac-38596

[9] B. Gammage, *The Broken Years: Australian Soldiers in the Great War.* Canberra: Australian National University Press, 1974.

[10] G. Simchoni, "Anne Frank's Diary: A Sentiment Analysis," 2017. [Online]. Available: http://giorasimchoni.com/2017/04/25/2017-04-25-anne-frank-s-diary-a-sentiment-analysis/

[11] C. Blevins, "Martha Ballard's Diary." [Online]. Available: http://historying.org/martha-ballards-diary/

[12] F. Boschetti, A. Cimino, F. Dell'orletta, G. E. Lebani, L. Passaro, P. Picchi, G. Venturi, S. Montemagni, and A. Lenci, "Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II," 2014.

[13] "Memorie Di Guerra." [Online]. Available: http://www.memoriediguerra.it/site

[14] A. Kuchling, "Regular Expression HOWTO," 2020. [Online]. Available: https://docs.python.org/3/howto/regex.html#regex-howto

[15] "re - Regular expression operations," 2020. [Online]. Available: https://docs.python.org/3/library/re.html

[16] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," in *Handbook of latent semantic analysis*, T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, Eds. Lawrence Erlbaum Associates Publishers, 2007, pp. 427–448.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[18] J. Beaumont, *Australia's War 1914-18.* St Leonards: Allen & Unwin Australia Pty Ltd, 1995.

[19] E. M. Andrews, *The Anzac Illusion: Anglo-Australian relations during World War I.* Cambridge: Cambridge University Press, 1993.

[20] The Australian War Memorial, "Enlistment statistics, First World War," 2018. [Online]. Available: https://www.awm.gov.au/articles/encyclopedia/enlistment/ww1

[21] ——, "Strong crowd numbers for Anzac Day Dawn Service - 2019," 2019. [Online]. Available: https://www.awm.gov.au/media/press-releases/DawnService2019

[22] State Library of New South Wales, "Personal diaries and letters from the First World War," 2019. [Online]. Available: https://www.sl.nsw.gov.au/research-and-collections/significant-collections/personal-diaries-and-letters-first-world-war

[23] ——, "Diarists from World War I," 2020. [Online]. Available: https://www.sl.nsw.gov.au/research-and-collections-significant-collections-world-war-i-collection/diarists-world-war-i

[24] JSON.org, "JSON." [Online]. Available: https://www.json.org/

[25] L. Richardson, "beautifulsoup4," 2020. [Online]. Available: https://pypi.org/project/beautifulsoup4/

[26] B. Rudis and B. Embrey, "pluralize: Pluralize and 'Singularize' Any (English) Word," 2020. [Online]. Available: https://cran.r-project.org/package=pluralize

[27] S. Thurner, R. Hanel, B. Liu, and B. Corominas-Murtra, "Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation," *Journal of the Royal Society, Interface*, vol. 12, 2014.

[28] Z. K. Silagadze, "Citations and the Zipf-Mandelbrot's law," *Complex Syst.*, vol. 11, 1997.

[29] The Python Software Foundation, "datetime Basic date and time types," 2020. [Online]. Available: https://docs.python.org/3/library/datetime.html

[30] D. Dufour, "date-extractor," 2020. [Online]. Available: https://pypi.org/project/date-extractor/

[31] G. Niemeyer, T. Pieviläinen, Y. de Leeuw, and P. Ganssle, "dateutil - powerful extensions to datetime," 2019. [Online]. Available: https://dateutil.readthedocs.io/en/stable/index.html

[32] Scrapinghub, "dateparser - python parser for human readable dates," 2014. [Online]. Available: https://dateparser.readthedocs.io/en/latest/index.html

[33] A. Koumjian and G. Corradini, "datefinder - extract dates from text," 2016. [Online]. Available: https://datefinder.readthedocs.io/en/latest/

[34] W. E. Hart, J.-P. Watson, and D. L. Woodruff, "Pyomo: modeling and solving mathematical programs in Python," *Mathematical Programming Computation*, vol. 3, no. 3, pp. 219–260, 2011.

[35] W. E. Hart, C. D. Laird, J.-P. Watson, D. L. Woodruff, G. A. Hackebeil, B. L. Nicholson, and J. D. Siirola, *Pyomo–optimization modeling in python*, 2nd ed. Springer Science & Business Media, 2017, vol. 67.

[36] L. Gurobi Optimization, "Gurobi Optimizer Reference Manual," 2020. [Online]. Available: https://www.gurobi.com/

[37] The Australian War Memorial, "Conscription, 1916-17." [Online]. Available: https://www.awm.gov.au/visit/exhibitions/anzac-voices/conscription

[38] Australian Electoral Commission, "Election dates 1901 present," 2020. [Online]. Available: https://www.aec.gov.au/elections/federal_elections/election-dates.htm

[39] J. K. Taubenberger and D. M. Morens, "1918 Influenza: The mother of all pandemics," *Emerging Infectious Diseases*, vol. 12, no. 1, pp. 15–22, 2006.

[40] S. Kuusk, *Fang farriers, Australian army dentistry in war and peace : a history of the Royal Australian Army Dental Corps. Volume 1, 1914-1939*, 2015.

[41] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," 2003.

[42] W. Zhang, T. Yoshida, and X. Tang, "TFIDF, LSI and Multi-word in Information Retrieval and Text Categorization," in *2008 IEEE International Conference on Systems, Man and Cybernetics*, 2008, pp. 108–113.

[43] D. Stevens, "November 1914 - Australia's First Victory at Sea," 2018. [Online]. Available: https://www.navy.gov.au/history/feature-histories/november-1914-australias-first-victory-sea

[44] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[45] D. Blei, "Probabilistic Topic Models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[46] T. Hofmann, "Probabilistic Latent Semantic Analysis," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 2013, pp. 289–296.

[47] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2017.

[48] J. Brownlee, "Overfitting and Underfitting With Machine Learning Algorithms," 2019. [Online]. Available: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

[49] G. Casella and E. George, "Explaining the Gibbs Sampler," *The American Statistician*, vol. 46, no. 3, pp. 167–174, 8 1992.

[50] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou, "A heuristic approach to determine an appropriate number of topics in topic modeling," *BMC Bioinformatics*, vol. 16, no. 13, 2015.

[51] R. Arun, V. Suresh, C. Veni Madhavan, and M. Narasimha Murty, "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations," in *Advances in Knowledge Discovery and Data Mining, Part I*, M. Zaki, J. Yu, B. Ravindran, and V. Pudi, Eds. Berlin. Heidelberg: Springer, 2010, pp. 391 – 402.

[52] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive LDA model selection," *Neurocomputing*, vol. 72, no. 7-9, pp. 1775 – 1781, 2008.

[53] R. Deveaud, E. SanJuan, and P. Bellot, "Accurate and effective Latent Concept Modeling for ad hoc information retrieval," *Document Numerique*, vol. 17, no. 1, pp. 61–84, 2014.

[54] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 5228–5235, 2004.

[55] M. Gerlach, T. P. Peixoto, and E. G. Altmann, "A network approach to topic models," *Science Advances*, vol. 4, no. 7, 7 2018.

[56] M. Nikita, "ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters," 2020. [Online]. Available: https://cran.r-project.org/package=ldatuning

[57] B. Grün and K. Hornik, "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software*, vol. 40, no. 13, pp. 1–30, 2011.

[58] R. Hyndman, "Moving averages," Tech. Rep., 2009.

[59] A. Zeileis and G. Grothendieck, "zoo: S3 Infrastructure for Regular and Irregular Time Series," *Journal of Statistical Software*, vol. 14, no. 6, pp. 1–27, 2005.

[60] The Australian War Memorial, "1918: Australians in France - Home at last - the Australians return." [Online]. Available: https://www.awm.gov.au/visit/exhibitions/1918/victory/returnhome

[61] DVA, "Repatriation of Australians in World War I," *DVA (Department of Veterans' Affairs) Anzac Portal*, 2020. [Online]. Available: https://anzacportal.dva.gov.au/wars-and-missions/ww1/politics/repatriation

[62] A. J. Reagan, C. M. Danforth, B. Tivnan, J. R. Williams, and P. S. Dodds, "Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs," *EPJ Data Science*, vol. 6, no. 1, 2017.

[63] T. Rinker, "lexicon: Lexicon Data," Buffalo, New York, 2018. [Online]. Available: http://github.com/trinker/lexicon

[64] F. . Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," in *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, vol. 718, 2011, pp. 93–98.

[65] M. M. Bradley and P. J. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings," Tech. Rep., 1999.

[66] T. Rinker, "Package 'lexicon'," 2019. [Online]. Available: https://cran.r-project.org/web/packages/lexicon/lexicon.pdf

[67] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings of the National Conference on Artificial Intelligence*, 2004, pp. 755–760. [Online]. Available: www.aaai.org

[68] T. Loughran and B. Mcdonald, "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, vol. 54, no. 4, pp. 1187–1230, 9 2016.

[69] S. M. Mohammad and P. D. Turney, "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon," in *Proceedings ofthe NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation ofEmotion in Text*, Los Angeles, 2010, pp. 26–34. [Online]. Available: http://www.wjh.harvard.edu/

[70] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives," in *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 2016, pp. 2666–2677.

[71] S. Baccianella, A. Esuli, and F. Sebastiani, "SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 2010, pp. 2200–2204. [Online]. Available: http://wordnetcode.princeton.

[72] C. Fellbaum, "WordNet(s)," in *Encyclopedia of Language and Linguistics*, 2nd ed., 2006, pp. 665–670.

[73] M. Jockers, "Syuzhet: Extract Sentiment and Plot Arcs from Text," 2015. [Online]. Available: https://github.com/mjockers/syuzhet

[74] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. 57, pp. 117–126, 9 2016.

[75] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Phildelphia, 2002, pp. 79–86.

[76] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *International Conference on Information and Knowledge Management*, Hong Kong, 2009, pp. 375–384.

[77] W. Francis and H. Kucera, "Brown Corpus Manual," 1971. [Online]. Available: http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM

[78] S. Evert and M. Baroni, "zipfR: Word Frequency Distributions in R," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, Prague, 2007.

[79] National Archives of Australia, "Abbreviations used in World War I and World War II service records," 2019. [Online]. Available: https://www.naa.gov.au/explore-collection/defence-and-war-service-records/researching-war-service/abbreviations-used-world-war-i-and-world-war-ii-service-records

Appendix A: Timeline of Australia in World War I

The following timeline is based on the book *Australia's War 1914-18* [18], and the Encyclopedia Britannica [1].

1914

28th June: Assassination of Austrian Archduke Franz Ferdinad

28th July: Austria declares war on Serbia

4th August: Britain declared war on Germany

6th August: Britain accepts Australia's offer of troops

11th - 17th September: Australian Naval and Military Expeditionary Force capture German New Guinea

1st November: First Contingent of the First Australian Imperial Force (AIF) leaves Australia

9th November: The Battle of the Cocos

3rd December: First AIF arrives in Egypt

1915

3rd February: Turkish attack on Suez Canal repelled

25th April: Landing at Gallipoli

6th - 8th August: Attack at Lone Pine

7th August: Attack at the Nek

19th - 20th December: Withdrawal from Gallipoli Peninsula

1916

14th March: First unit of the Australian Flying Corps prepares to leave for the front

1st July: Battle of the Somme begins

19th - 20th July: Battle of Fromelles

23rd July: Attempt to capture Pozieres begins

15th September: First use of tanks on the Somme

28th October: First Conscription Referendum

1917

6th April: United States declares war on Germany

11th April: First battle of Bullecourt

3rd May: Second battle of Bullecourt

5th May: Australian 1917 Federal Election

7th June: Battle of Messines begins

31st July: Third battle of Ypres begins

20th September: Battle of Menin Road

26th September: Battle of Polygon Wood

4th October: Attack on Broodseinde Ridge

12th October: Attack on town of Passchendaele

31st October: Charge of Light Horse at Beersheba

29th November: Prime Minister Billy Hughes pelted with egg

9th December: British troops capture Jerusalem

20th December: Second Conscription Referendum

1918

21st March: German offensive at Arras

4th July: Australian troops capture Hamel

8th August: Allied offensive at Amiens

18th September: Attack on the Hindenburg line

1st October: Damascus captured

11th November: Armistice is signed with Germany at 11am

Appendix B: World War I Casualties

Table B.1 contains approximate military casualty figures from World War I as given in the Encyclopædia Britannica [1].

Country	Total Mobilised Forces	Deaths	Wounded	Prisoners and Missing	Total Casualties	Percentage of Mobilized Forces in Casualties
Allied and Associated Powers						
Russia	12,000,000	1,700,000	4,950,000	2,500,000	9,150,000	76.3%
British Empire	8,904,467	908,371	2,090,212	191,652	3,190,235	35.8%
France	8,410,000	1,357,800	4,266,000	537,000	6,160,800	73.3%
Italy	5,615,000	650,000	947,000	600,000	2,197,000	39.1%
United States	4,355,000	116,516	204,002	4,500	323,018	8.1%
Japan	800,000	300	907	3	1,210	0.2%
Romania	750,000	335,706	120,000	80,000	535,706	71.4%
Serbia	707,343	45,000	133,148	152,958	331,106	46.8%
Belgium	267,000	13,716	44,686	34,659	93,061	34.9%
Greece	230,000	5,000	21,000	1,000	27,000	11.7%
Portugal	100,000	7,222	13,751	12,318	33,291	33.3%
Montenegro	50,000	3,000	10,000	7,000	20,000	40.0%
Total	42,188,810	5,142,631	12,800,706	4,121,090	22,064,427	52.3%
Central Powers						
Germany	11,000,000	1,773,700	4,216,058	1,152,800	7,142,558	64.9%
Austria-Hungary	7,800,000	1,200,000	3,620,000	2,200,000	7,020,000	90.0%
Turkey	2,850,000	325,000	400,000	250,000	975,000	34.2%
Bulgaria	1,200,000	87,500	152,390	277,029	266,919	22.2%
Total	22,850,000	3,386,200	8,388,448	3,629,829	15,404,477	67.4%
Grand Total	**65,038,810**	**8,58,831**	**21,189,154**	**7,750,919**	**37,468,904**	**57.5%**

Table B.1: Approximate military casualties in World War I, sourced from the U.S. War Department with U.S. casualties as amended by the Statistical Services Center, Office of the Secretary of Defence. Reprinted from the Encyclopædia Britannica [1].

Appendix C: JSON Page File


```
{
  "field_status": {
    "value": null,
    "workflow": null
  },
  "body": {
    "value": "<p class=\"page\" id=\"a3038010\">[Page 10]</p>\n<p>1917<br />Sat. 2
        nd June.<br />I returned to my interrupted tea still rather suspicious
        taking the precaution of closing switches and preparing for transmission
        from the accumulators, since the dynamo was not then running.  No signs of
        life in the 'phones.  At 4.45 pm. there was a rush on deck     a 'plane
        motor could be heard just outside and I got on deck just in time to see a
        bomb drop off the port bow and a message attached to a sandbag fall on deck
         from a seaplane just overhead.  The machine was a small two-seater biplane
             a German Naval Ensign was clearly visible flying from one of her stays
             and she was flying so low that her pilot and observer could be
        distinctly seen.  The latter was apparently ready to do some serious
        bombing.  I made for the bridge for orders, but before I reached the
        companionway instructions came down \"not to use the wireless\"!</p>\n",
    "summary": "",
    "format": "filtered_html"
  },
  "field_document": {
    "uri": "https://transcripts.sl.nsw.gov.au/node/80404",
    "id": "80404",
    "resource": "node",
    "revision_id": "80418",
    "uuid": "1a679885-fc1a-4d70-9453-c4f5b37ad6b9"
  },
  "field_number": "10",
  "field_diary_title": "Alexander diary, 1917-1918 / Roy Alexander",
  "field_transcript": [

  ],
  "field_working_notes": null,
  "field_review_notes": null,
  "field_tt_image": {
    "file": {
      "uri": "https://transcripts.sl.nsw.gov.au/file/100623",
      "id": "100623",
      "resource": "file",
      "uuid": "28be9771-f1a3-4801-bf66-b592beba3d66"
    },
    "alt": ""
```

```
  },
  "field_canonical_id": "WW1 Document MLMSS 1610 p.010",
  "nid": "103910",
  "vid": "906499",
  "is_new": false,
  "type": "page",
  "title": "Alexander diary, 1917-1918 / Roy Alexander - Page 10",
  "language": "und",
  "url": "https://transcripts.sl.nsw.gov.au/page/alexander-diary-1917-1918-roy-
      alexander-page-10",
  "edit_url": "https://transcripts.sl.nsw.gov.au/node/103910/edit",
  "status": "1",
  "promote": "0",
  "sticky": "0",
  "created": "1434774345",
  "changed": "1435378140",
  "author": {
    "uri": "https://transcripts.sl.nsw.gov.au/user/1",
    "id": "1",
    "resource": "user",
    "uuid": "a5ca06b0-1bb9-4483-aa03-cea5b7cdd46a"
  },
  "log": "",
  "revision": null,
  "comment": "1",
  "comments": [

  ],
  "comment_count": "0",
  "comment_count_new": "0",
  "uuid": "f5e1eb2e-d33e-4bcb-90ba-d7af2c1efedd",
  "vuuid": "8d32ff43-b734-4bab-a4b0-cdb404461f9e",
  "metatag": {
    "value": {
      "image_src": "https://prod--slnsw-transcription-tool.s3.ap-southeast-2.
          amazonaws.com/s3fs-public/files/transcript_image/a3038010h_0.jpg",
      "transcriptiontool_hierarchylevel": "Page",
      "transcriptiontool_canonicalid": "WW1 Document MLMSS 1610 p.010",
      "transcriptiontool_collection": "WW1",
      "transcriptiontool_section": "WW1 Series World War 1 Diaries",
      "transcriptiontool_document": "WW1 Document MLMSS 1610",
      "transcriptiontool_page": "10",
      "title": "Alexander diary, 1917-1918 / Roy Alexander - Page 10 |
          Transcription Tool",
      "description": "[Page 10]",
      "canonical_url": "https://transcripts.sl.nsw.gov.au/page/alexander-diary-1917
          -1918-roy-alexander-page-10",
      "shortlink": "https://transcripts.sl.nsw.gov.au/node/103910",
      "generator": "Drupal 7 (https://www.drupal.org)"
    }
  }
}
```


Listing C.1: JSON page file from Roy Alexander's Diary.

Appendix D: Common Abbreviations and Stop Words

Table D.1 contains a list of words and their abbreviations. In order to clean the diaries it was required that these abbreviations were converted to the full word so that they would be counted as the same word. Some of these abbreviations were found when running our initial analysis on the diaries, whilst others were known based on the National Archives of Australia's list of abbreviations used in service records [79]. For instance, *lre* was found when initially running our tf-idf analysis, and through close reading of the diaries was found to mean letter, such as in Leo Walsh's diary:

> "Mail Day. Lres. from Roseby, Miss Vera Brown, Miss Warby, Buck & photos, 2 from Ede, most of these been to Lake Michigan." and "Mail day. Sent lres. to Ede & Edna."

We note that not all abbreviations could be included as they had more than one meaning, for example: *brig* could mean brigade or brigadier or refer to a type of ship, *pt* is short for port and for the Egyptian currency *piastre*, and *lieu* is both a word in its own right and short for lieutenant.

In our data cleaning we remove stop words based on the `stop_words` data set in the `tidytext` package [2] in R. We also add our own stop words to this list and these are given in Table D.2.

Word	**Abbreviations**	**Word**	**Abbreviations**
about	abt	hospital	hosp
afternoon	aftn	infantry	inf, infy
arrive	arr	letter	lre
artillery	arty	lieutenant	lieut, lt
australia	aus, aust	major	maj
battalion	bn, batt, battn	melbourne	melb, melbne
battery	bty	morning	morn
brigade	bde	packet	pkt
british	btsh	pioneer	pnr
captain	capt	private	pte, pvt
christmas	xma	railway	rwy
colonel	col	received	rec, recd
company	coy	regiment	regt
corporal	cpl	reinforcement	reinf
department	dept	sergeant	sgt, sergt
division	div	signal	sig
dozen	doz	squadron	sqn
engineer	engr	surgeon	surg, surgn
general	gen	yesterday	yesty
headquater	hq, hqr, hdqr		

Table D.1: Common abbreviations found in the World War I diaries. These abbreviations had to be converted back to the full word before analysis so that they would be counted as the same word.

monday	lun	feby	nov	juillet
mon	mardi	march	december	juill
tuesday	mar	april	decbr	juil
tues	mercredi	apr	dec	aout
tue	mer	apl	janvier	août
wednesday	jeudi	may	janv	septembre
wed	jeu	june	fevrier	octobre
thursday	vendredi	july	février	novembre
thurs	ven	august	fevr	decembre
thu	samedi	aug	févr	décembre
friday	sam	september	fév	déc
fri	dimanche	septr	mars	st
saturday	dim	sept	mar	th
sat	january	sep	avril	rd
sunday	jan	october	avr	nd
sun	february	oct	mai	html
lundi	feb	november	juin	amp

Table D.2: These words were manually added to the `stop_words` data set from the `tidytext` package [2] in R to form our full set of stop words for this thesis.

Appendix E: Examples of Dates from the Diaires

Some examples of the ways dates are written in the diaries are given below. Note that the dates are highlighted in blue, whilst items that resemble dates but are not are given in red.

- **Document 2:** "1917 Sat. 2nd June Weather perfect. Dead calm. We ... "
- **Document 3:** "7th Arrived at the Island of Lemnos, not much of ... "
- **Document 6:** "21st. Jan, Church Parade this morning & spell this ... 22nd. & 23rd. Bull Ring Parade both days. ... "
- **Document 15:** "Saturday, April 8th: Company and rifle drill in morning. ... April 9th: Early morning parade, also Church parade. ... "
- **Document 16:** "Saturday 19/12/15. It is very funny to see the fellows ..."
- **Document 20:** "Saturday April 21 Tom gone to Sydney for week ... "
- **Document 23:** "Sunday, 10th June. Weather squally with a moderate ... "
- **Document 25:** "April 1915 1 Thursday Saw B Sqn AHQ swim their horses Gardens ... 2 Fri Rode back with Clark ... "
- **Document 26:** "14th Friday : Self & Lt Lindsell go to Serapeum in ... "
- **Document 43:** "On Monday, 23rd October 1916, I embarked from ... "
- **Document 47:** "Oct. 21st. We passed Zyra at 1 p.m. today, also a destroyer ... 22nd we passed Sicily close in to shore, but owing to a thick haze ... "
- **Document 53:** "Monday 4th Decr Got up about 7 oclock and had ... "

- **Document 64:** "August 26th Went to Rest Gully again and relieved the 8th and 10th Light Horse on top of Walkers Ridge. . . . "
- **Document 76:** "3-6-16 Trekked at 4 a.m. for 8 miles camped at . . . "
- **Document 89:** "17.4.15 Sill steaming along the Gulf of Suez this . . . "
- **Document 119:** "Jan 10/16 The "BILLY'S" have arrived, for months . . . "
- **Document 129:** "Dec 7/ Left Mex camp at 11.30 pm My old horse Kaisa Bill gave me . . . 8th/ Arrived at Railhead at 10-3 am after . . . "
- **Document 142:** "Jany 1915 Scare 15th At 9.10 am Ulysses sent signal . . . "
- **Document 524:** "Janvier 19 Dimanche Went to Orchestra Symphony Concert at the Alhambra Theatre. . . . 20 Lundi Getting on well. It is touch . . . "
- **Document 538:** "22nd October (Sunday) Up about 8.30 am after . . . "
- **Document 620:** "Mond. 5th March 1917 Yesterday morning we had . . . "
- **Document 621:** "Friday 25th Jany 18 Left London for Edinburgh 10/15 a.m. arriving . . . Saty 2nd. Left Aberdeen Fri evening at 8/15 p.m. and arrived . . . 1918 Feby 8th to 27th. Quiet period. . . . "
- **Document 635:** "Decembre 23 We are doing a fair amount of Section . . . "
- **Document 719:** "7:5:18 Roused from my bed in haste this morning by . . . "
- **Document 777:** "28. Thursday Stayed in my room without fire caught . . . "
- **Document 834:** "February 1915 (Thursday 11th) Left Liverpool camp at . . . (Friday 12th) At sea good few of the boys are sea sick still, still on . . . "
- **Document 862:** "12 Th April Captured one of the Turkish Lancers . . . 13 Fri April Very quiet Heavy artillery bombardment and Fokkers very . . . "
- **Document 936:** "Jan': 20 Gulf of Aden arrive at Aden; what . . . "

Appendix F: Topics

Based on Figure 5.16, ten topics were used in our LDA analysis. Tables F.1 - F.10 give the top 99 most probable terms in each topic. We have given each topic a name based on what the majority of the top 99 words within the topic were referring to, and in some cases where this topic peaked in Figure 5.17. Our topic names are *Everyday life*, *War at Sea*, *Egypt*, *Gallipoli*, *In the Trenches (Beginning)*, *In the Trenches (Middle)*, *In the Trenches (End)*, *White Christmas*, *After the Armistice*, and *Home Again*. Below we give brief explanations regarding the name of each topic, and some of the most interesting words from that topic.

Everyday Life (Table F.1):

In this topic we see many everyday words, such as words relating to the time of day, meal times, military ranks, places (*camp*, *church*, *hospital*), and transport (*arrived*, *ship*, *train*). None of these words relate to specific battles or events, hence, this topic was called *Everyday Life*.

War at Sea (Table F.2):

In this topic we see words regarding several types of ships, as well as the names of at least two ships: the *Emden* and the *Berrima*. We also see place names around New Guinea, such as *Rabaul* and *Herbertshohe*. These words are all related to the Australian occupation of German New Guinea, and hence this topic has been name *War at Sea*.

Egypt (Table F.3):

In this topic we see the names of several places in Egypt, including *Cairo*, *Alexandria*, the *Suez Canal*, the *Nile*, *Ismailia*, *Heliopolis* and *Tel el Kebir*. We also see words such as *pyramid*, *camel*, *sand*, *desert*, *piastre*, *training* and *marching*. All of these words are related to the time spent by Australian soldiers whilst training in

Egypt. Hence, this topic was called *Egypt*.

Gallipoli (Table F.4):

In this topic we see general words regarding fighting, including *artillery*, *wounded*, *trench* and *infantry*. However, we also see several words specific to the fighting at Gallipoli such as *turk*, *turkish*, *beach*, *landed*, and *submarine*. Locations from this campaign are also present, including *Lemnos*, *Gallipoli*, *Port Mudros* and the *Dardanelles*. We do still see some place names from Egypt, such as *Cairo* and *Alexandria*, however, since the majority of words relate to the Gallipoli campaign this topic was called *Gallipoli*.

In the Trenches (Beginning), (Middle), and (End). (Table F.5, Table F.6, Table F.7):

In all three of these topics we see general words related to fighting such as *artillery*, *wounded*, *shelling*, *gas* and *infantry*. They also all include the word *fritz*, with Table F.5 and Table F.7 also including the words *hun* and *German*. Both *fritz* and *hun* were nicknames for German soldiers, and hence these terms suggest these topics are about the fighting on the Western Front in Europe. This is further supported by the appearance of locations throughout Europe including *France*, *Amiens*, *Bapaume*, *Ypres*, the *Somme* and *Peronne*. In Table F.6 and Table F.7 we also see locations within the Middle East including *Gaza*, *Jerusalem*, *Jordan* and *Jericho*. Since these three topics all contain similar words regarding the battles on the Western Front and in the Middle East they are instead differentiated by when they are most prominent in Figure 5.17. Hence, the topics given in Table F.5, Table F.6 and Table F.7 were named *In the Trenches (Beginning)*, *In the Trenches (Middle)*, and *In the Trenches (End)*, respectively.

White Christmas (Table F.8):

In this topic we see many words regarding cold weather, including *snow*, *frost*, *fog*, *ice*, *freezing*, *wet* and *sleet*. We also see words associated with Christmas, including *Christmas*, *parcel*, *rum* and *pudding*. Hence, this topic was called *White Christmas*. Interestingly, this topic also includes the word *miserable*, suggesting that the cold and rain made the conditions in the trenches a horrible place causing many to feel miserable. The words *conscription* and *vote* also appear in this topic. This would be because the 1916 and 1917 conscription referendums happened in October and

December, respectively, which coincides with winter in Europe.

After the Armistice (Table F.9):

In this topic we do not see any words regarding battles and instead see words such as *peace*, *armistice* and *civilian*. We also see words regarding leisure activities such as *walk*, *dance*, *concert*, *theatre*, *read* and *visit*, and locations around Europe, including *London*, *Paris* and *Charleroi*. This suggests that this topic is about the time after the armistice was signed when soldiers took trips to various cities around Europe whilst waiting for transport back to Australia. Hence, this topic was called *After the Armistice*.

Home Again (Table F.10):

This topic does not include any words regarding battles, instead we see words such as *home*, *bed*, *arrive*, *shopping*, *mum*, *dad* and *auntie*. These words all suggest that this topic is about when they are back home with their families. Hence, this topic was named *Home Again*.

Everyday Life

rank	term	beta	rank	term	beta	rank	term	beta
1	day	0.0213	34	bed	0.0024	67	tomorrow	0.0018
2	night	0.0132	35	company	0.0024	68	hand	0.0018
3	morning	0.0115	36	parade	0.0024	69	bad	0.0018
4	time	0.0095	37	battalion	0.0023	70	sleep	0.0017
5	left	0.0070	38	light	0.0023	71	german	0.0017
6	afternoon	0.0068	39	troop	0.0022	72	person	0.0017
7	pm	0.0057	40	station	0.0022	73	rest	0.0017
8	camp	0.0051	41	horse	0.0022	74	wrote	0.0017
9	letter	0.0047	42	australian	0.0022	75	australia	0.0017
10	arrived	0.0047	43	french	0.0021	76	half	0.0017
11	mile	0.0042	44	boy	0.0021	77	head	0.0017
12	home	0.0042	45	duty	0.0021	78	captain	0.0017
13	tea	0.0042	46	returned	0.0021	79	major	0.0017
14	hour	0.0042	47	hot	0.0021	80	ground	0.0017
15	round	0.0039	48	heavy	0.0021	81	house	0.0017
16	officer	0.0038	49	train	0.0020	82	hill	0.0016
17	line	0.0037	50	war	0.0020	83	guard	0.0016
18	leave	0.0035	51	breakfast	0.0020	84	started	0.0016
19	fine	0.0033	52	church	0.0020	85	sea	0.0016
20	evening	0.0032	53	tonight	0.0019	86	foot	0.0016
21	till	0.0032	54	colonel	0.0019	87	fire	0.0016
22	dinner	0.0031	55	party	0.0019	88	post	0.0016
23	water	0.0030	56	yesterday	0.0019	89	bit	0.0016
24	received	0.0029	57	beautiful	0.0019	90	ship	0.0016
25	weather	0.0027	58	called	0.0019	91	spent	0.0015
26	brigade	0.0027	59	rain	0.0019	92	close	0.0015
27	hospital	0.0026	60	told	0.0018	93	gun	0.0015
28	found	0.0025	61	pretty	0.0018	94	couple	0.0015
29	town	0.0025	62	week	0.0018	95	mail	0.0015
30	usual	0.0025	63	road	0.0018	96	life	0.0014
31	lot	0.0025	64	coming	0.0018	97	front	0.0014
32	passed	0.0025	65	sergeant	0.0018	98	hard	0.0014
33	cold	0.0024	66	met	0.0018	99	move	0.0014

Table F.1: Top 99 terms for the *Everyday Life* topic with their probabilities.

War at Sea

rank	term	beta	rank	term	beta	rank	term	beta
1	ship	0.0157	34	steamer	0.0026	67	anchored	0.0017
2	sydney	0.0098	35	cruiser	0.0026	68	passed	0.0017
3	german	0.0092	36	colombo	0.0024	69	health	0.0017
4	officer	0.0076	37	berrima	0.0024	70	official	0.0016
5	captain	0.0074	38	crew	0.0024	71	governor	0.0016
6	boat	0.0068	39	gun	0.0023	72	knot	0.0016
7	lieutenant	0.0066	40	australia	0.0023	73	action	0.0016
8	board	0.0063	41	herbertshohe	0.0023	74	clock	0.0016
9	island	0.0060	42	ashore	0.0022	75	alongside	0.0016
10	troop	0.0058	43	military	0.0022	76	navy	0.0016
11	native	0.0054	44	drill	0.0022	77	admiral	0.0015
12	received	0.0047	45	arrived	0.0022	78	landed	0.0015
13	message	0.0047	46	report	0.0021	79	satisfactory	0.0015
14	wireless	0.0046	47	commander	0.0021	80	signalling	0.0015
15	sea	0.0046	48	holme	0.0021	81	guard	0.0015
16	colonel	0.0045	49	signal	0.0020	82	expedition	0.0015
17	rabaul	0.0045	50	prisoner	0.0020	83	guinea	0.0015
18	port	0.0044	51	convoy	0.0019	84	shore	0.0015
19	deck	0.0042	52	anchor	0.0019	85	wharf	0.0015
20	company	0.0040	53	british	0.0019	86	aboard	0.0014
21	harbour	0.0038	54	brigadier	0.0019	87	albany	0.0014
22	emden	0.0038	55	coal	0.0019	88	duty	0.0014
23	naval	0.0037	56	returned	0.0019	89	land	0.0014
24	horse	0.0036	57	instruction	0.0019	90	collier	0.0014
25	administrator	0.0034	58	aden	0.0019	91	return	0.0014
26	melbourne	0.0031	59	warship	0.0018	92	police	0.0014
27	government	0.0030	60	encounter	0.0018	93	water	0.0014
28	fleet	0.0029	61	bay	0.0018	94	flagship	0.0014
29	force	0.0029	62	vessel	0.0018	95	war	0.0014
30	major	0.0028	63	infantry	0.0017	96	reported	0.0014
31	station	0.0028	64	command	0.0017	97	eastern	0.0014
32	flag	0.0028	65	service	0.0017	98	proceed	0.0013
33	garrison	0.0027	66	store	0.0017	99	private	0.0013

Table F.2: Top 99 terms for the *War at Sea* topic with their probabilities.

Egypt

rank	term	beta	rank	term	beta	rank	term	beta
1	cairo	0.0171	34	passed	0.0026	67	lh	0.0015
2	canal	0.0128	35	land	0.0025	68	lecture	0.0014
3	parade	0.0099	36	serapeum	0.0023	69	marseille	0.0014
4	camp	0.0092	37	colombo	0.0023	70	guard	0.0014
5	sand	0.0085	38	trench	0.0022	71	picket	0.0014
6	horse	0.0085	39	route	0.0022	72	orderly	0.0014
7	ship	0.0083	40	ashore	0.0022	73	ride	0.0014
8	tent	0.0080	41	squadron	0.0022	74	money	0.0014
9	desert	0.0079	42	dust	0.0022	75	rifle	0.0014
10	native	0.0070	43	mosque	0.0022	76	oclock	0.0013
11	el	0.0065	44	marching	0.0021	77	practice	0.0013
12	drill	0.0062	45	donkey	0.0021	78	ferry	0.0013
13	egypt	0.0051	46	maadi	0.0021	79	fellow	0.0013
14	troop	0.0051	47	fatigue	0.0020	80	cool	0.0013
15	sea	0.0047	48	piastre	0.0020	81	gallipoli	0.0013
16	camel	0.0044	49	heat	0.0020	82	musketry	0.0013
17	regiment	0.0044	50	soldier	0.0019	83	colonel	0.0013
18	egyptian	0.0044	51	fuller	0.0019	84	board	0.0013
19	tel	0.0043	52	revielle	0.0019	85	stable	0.0013
20	train	0.0042	53	nigger	0.0019	86	training	0.0013
21	kebir	0.0039	54	bag	0.0018	87	railway	0.0013
22	alexandria	0.0038	55	artillery	0.0018	88	hot	0.0013
23	pyramid	0.0038	56	arab	0.0018	89	mena	0.0012
24	heliopoli	0.0037	57	wharf	0.0018	90	tucker	0.0012
25	deck	0.0036	58	signalling	0.0018	91	christmas	0.0012
26	boat	0.0036	59	indian	0.0017	92	about	0.0012
27	island	0.0035	60	infantry	0.0017	93	citadel	0.0012
28	suez	0.0034	61	reinforcement	0.0016	94	calm	0.0012
29	water	0.0033	62	class	0.0016	95	cox	0.0012
30	harbour	0.0029	63	mounted	0.0015	96	fort	0.0012
31	sydney	0.0028	64	signaller	0.0015	97	swimming	0.0012
32	nile	0.0028	65	major	0.0015	98	riding	0.0012
33	ismailia	0.0027	66	port	0.0015	99	waggon	0.0012

Table F.3: Top 99 terms for the *Egypt* topic with their probabilities.

Gallipoli

rank	term	beta	rank	term	beta	rank	term	beta
1	turk	0.0194	34	lemno	0.0031	67	burst	0.0020
2	trench	0.0185	35	landing	0.0030	68	chap	0.0020
3	gun	0.0131	36	alexandria	0.0029	69	ammunition	0.0020
4	shell	0.0123	37	casualty	0.0029	70	indian	0.0020
5	wounded	0.0097	38	water	0.0028	71	howitzer	0.0020
6	ship	0.0094	39	dug	0.0028	72	deck	0.0019
7	fire	0.0085	40	warship	0.0028	73	gallipoli	0.0019
8	enemy	0.0081	41	anzac	0.0027	74	dugout	0.0019
9	firing	0.0075	42	landed	0.0027	75	head	0.0018
10	beach	0.0073	43	sniper	0.0027	76	mudro	0.0018
11	boat	0.0065	44	hit	0.0027	77	cape	0.0018
12	position	0.0062	45	board	0.0026	78	destroyer	0.0018
13	attack	0.0061	46	troop	0.0026	79	infantry	0.0018
14	shrapnel	0.0060	47	fired	0.0026	80	horse	0.0018
15	bomb	0.0056	48	damage	0.0025	81	fatigue	0.0018
16	battery	0.0054	49	machine	0.0025	82	dardanelle	0.0018
17	artillery	0.0053	50	dead	0.0025	83	flank	0.0017
18	sea	0.0050	51	aeroplane	0.0025	84	sick	0.0017
19	quiet	0.0048	52	front	0.0024	85	store	0.0016
20	hospital	0.0047	53	submarine	0.0024	86	arm	0.0016
21	turkish	0.0047	54	pm	0.0023	87	captain	0.0016
22	cairo	0.0045	55	harbour	0.0023	88	patient	0.0016
23	killed	0.0042	56	reinforcement	0.0023	89	mule	0.0016
24	rifle	0.0042	57	swim	0.0023	90	relieved	0.0016
25	hill	0.0040	58	aboard	0.0022	91	tepe	0.0016
26	shot	0.0040	59	shelling	0.0022	92	charge	0.0016
27	line	0.0038	60	ridge	0.0022	93	shelled	0.0016
28	bullet	0.0038	61	land	0.0021	94	sap	0.0016
29	bombardment	0.0036	62	transport	0.0021	95	cruiser	0.0015
30	ashore	0.0034	63	night	0.0021	96	colonel	0.0015
31	heavy	0.0033	64	oclock	0.0021	97	mail	0.0015
32	gully	0.0032	65	bay	0.0021	98	navy	0.0015
33	island	0.0031	66	shore	0.0021	99	greek	0.0015

Table F.4: Top 99 terms for the *Gallipoli* topic with their probabilities.

In the Trenches (Beginning)

rank	term	beta	rank	term	beta	rank	term	beta
1	shell	0.0160	34	position	0.0037	67	railway	0.0022
2	trench	0.0152	35	machine	0.0036	68	nice	0.0022
3	gun	0.0141	36	france	0.0036	69	walk	0.0022
4	line	0.0124	37	london	0.0035	70	infantry	0.0022
5	fritz	0.0101	38	killed	0.0034	71	route	0.0021
6	german	0.0083	39	stunt	0.0034	72	farm	0.0021
7	front	0.0072	40	fatigue	0.0033	73	quiet	0.0021
8	artillery	0.0071	41	oclock	0.0033	74	franc	0.0021
9	wounded	0.0069	42	fire	0.0033	75	walked	0.0021
10	billet	0.0069	43	hut	0.0032	76	boy	0.0021
11	gas	0.0066	44	drill	0.0032	77	somme	0.0021
12	bomb	0.0064	45	shelling	0.0031	78	shrapnel	0.0021
13	plane	0.0058	46	battalion	0.0031	79	prisoner	0.0020
14	mile	0.0058	47	el	0.0031	80	tent	0.0020
15	camp	0.0057	48	raining	0.0029	81	balloon	0.0019
16	marched	0.0055	49	train	0.0029	82	chap	0.0019
17	bombardment	0.0054	50	move	0.0028	83	hit	0.0019
18	village	0.0054	51	dug	0.0028	84	reveille	0.0018
19	heavy	0.0054	52	wood	0.0027	85	regiment	0.0018
20	road	0.0051	53	helmet	0.0027	86	amien	0.0017
21	firing	0.0050	54	town	0.0027	87	breakfast	0.0017
22	battery	0.0049	55	bapaume	0.0027	88	ypre	0.0017
23	horse	0.0045	56	wet	0.0026	89	turk	0.0017
24	casualty	0.0045	57	church	0.0025	90	bearer	0.0017
25	attack	0.0045	58	hun	0.0025	91	moved	0.0017
26	enemy	0.0041	59	raid	0.0025	92	field	0.0017
27	parade	0.0041	60	relieved	0.0024	93	england	0.0016
28	aeroplane	0.0039	61	dugout	0.0023	94	ammunition	0.0016
29	fine	0.0039	62	anzac	0.0023	95	ambulance	0.0016
30	evening	0.0038	63	shelled	0.0023	96	private	0.0016
31	division	0.0037	64	rain	0.0022	97	marching	0.0016
32	albert	0.0037	65	camel	0.0022	98	bir	0.0016
33	taube	0.0037	66	tommy	0.0022	99	truck	0.0015

Table F.5: Top 99 terms for the *In the Trenches (Beginning)* topic with their probabilities.

In the Trenches (Middle)

rank	term	beta	rank	term	beta	rank	term	beta
1	road	0.0069	34	machine	0.0025	67	christmas	0.0017
2	gun	0.0060	35	front	0.0023	68	heavy	0.0017
3	fine	0.0060	36	shelling	0.0023	69	weather	0.0017
4	fritz	0.0055	37	stunt	0.0023	70	shelled	0.0017
5	wrote	0.0052	38	letter	0.0023	71	wood	0.0016
6	ypre	0.0051	39	wounded	0.0022	72	franc	0.0016
7	line	0.0048	40	blighty	0.0022	73	jaffa	0.0016
8	enemy	0.0048	41	de	0.0022	74	hazebrouck	0.0016
9	brigade	0.0047	42	raining	0.0022	75	rested	0.0016
10	dinner	0.0043	43	gas	0.0022	76	gaza	0.0016
11	train	0.0040	44	lunch	0.0022	77	track	0.0015
12	bomb	0.0040	45	ridge	0.0022	78	poperinghe	0.0015
13	shell	0.0038	46	dump	0.0022	79	italian	0.0015
14	hut	0.0038	47	section	0.0022	80	played	0.0015
15	cold	0.0037	48	deferred	0.0022	81	tea	0.0015
16	bailleul	0.0033	49	lovely	0.0021	82	battery	0.0015
17	fed	0.0033	50	wet	0.0021	83	le	0.0015
18	lorry	0.0033	51	omer	0.0021	84	frost	0.0015
19	plane	0.0032	52	cleaned	0.0020	85	dropped	0.0015
20	raid	0.0032	53	village	0.0019	86	kilo	0.0015
21	sister	0.0030	54	rain	0.0019	87	bombing	0.0014
22	boulogne	0.0030	55	miss	0.0019	88	pay	0.0014
23	camel	0.0029	56	billet	0.0019	89	car	0.0014
24	report	0.0029	57	position	0.0019	90	walk	0.0014
25	division	0.0029	58	killed	0.0019	91	hotel	0.0013
26	barrage	0.0029	59	dugout	0.0019	92	eglise	0.0013
27	london	0.0027	60	mud	0.0019	93	wh	0.0013
28	walked	0.0026	61	spent	0.0019	94	boche	0.0013
29	farm	0.0026	62	box	0.0019	95	wolf	0.0013
30	pt	0.0026	63	moved	0.0018	96	jerusalem	0.0013
31	paris	0.0025	64	book	0.0018	97	station	0.0013
32	battalion	0.0025	65	pill	0.0017	98	artillery	0.0013
33	messine	0.0025	66	sector	0.0017	99	corner	0.0013

Table F.6: Top 99 terms for the *In the Trenches (Middle)* topic with their probabilities.

In the Trenches (End)

rank	term	beta	rank	term	beta	rank	term	beta
1	line	0.0139	34	captured	0.0030	67	jordan	0.0018
2	fritz	0.0134	35	evening	0.0029	68	tommy	0.0018
3	gun	0.0128	36	barrage	0.0028	69	corbie	0.0018
4	enemy	0.0102	37	move	0.0028	70	hit	0.0017
5	shell	0.0099	38	le	0.0027	71	bridge	0.0017
6	front	0.0085	39	la	0.0027	72	night	0.0017
7	village	0.0071	40	artillery	0.0026	73	dead	0.0017
8	plane	0.0067	41	casualty	0.0025	74	dressing	0.0017
9	road	0.0066	42	shelled	0.0025	75	load	0.0017
10	battalion	0.0061	43	american	0.0025	76	horse	0.0017
11	prisoner	0.0058	44	lorry	0.0025	77	raining	0.0016
12	hun	0.0057	45	advance	0.0024	78	limber	0.0016
13	bomb	0.0055	46	river	0.0024	79	girl	0.0016
14	division	0.0051	47	tank	0.0024	80	convoy	0.0016
15	attack	0.0048	48	dug	0.0023	81	raid	0.0016
16	wounded	0.0046	49	wood	0.0023	82	killed	0.0016
17	position	0.0045	50	air	0.0023	83	sector	0.0016
18	fine	0.0044	51	walked	0.0022	84	troop	0.0015
19	amien	0.0043	52	fire	0.0022	85	fighting	0.0015
20	battery	0.0043	53	aussie	0.0022	86	heavily	0.0015
21	somme	0.0041	54	chap	0.0022	87	attacked	0.0015
22	quiet	0.0040	55	brigade	0.0022	88	dropped	0.0015
23	trench	0.0039	56	kilo	0.0021	89	relieved	0.0015
24	stunt	0.0039	57	valley	0.0021	90	corps	0.0015
25	gas	0.0037	58	dump	0.0021	91	engine	0.0015
26	machine	0.0037	59	weather	0.0020	92	balloon	0.0015
27	moved	0.0036	60	peronne	0.0020	93	jericho	0.0015
28	shelling	0.0035	61	news	0.0019	94	waggon	0.0014
29	viller	0.0032	62	train	0.0019	95	objective	0.0014
30	heavy	0.0031	63	boche	0.0019	96	chateau	0.0014
31	french	0.0031	64	german	0.0019	97	yank	0.0014
32	dugout	0.0030	65	bombing	0.0018	98	bank	0.0014
33	forward	0.0030	66	park	0.0018	99	swim	0.0014

Table F.7: Top 99 terms for the *In the Trenches (End)* topic with their probabilities.

White Christmas

rank	term	beta	rank	term	beta	rank	term	beta
1	cold	0.0420	34	wet	0.0024	67	sand	0.0015
2	snow	0.0252	35	taube	0.0024	68	deep	0.0015
3	mud	0.0145	36	fall	0.0023	69	somme	0.0015
4	christmas	0.0124	37	ration	0.0022	70	boot	0.0015
5	hut	0.0077	38	miserable	0.0022	71	cleaning	0.0014
6	fritz	0.0073	39	blighty	0.0022	72	comfort	0.0014
7	el	0.0070	40	snowed	0.0022	73	conscription	0.0014
8	frozen	0.0068	41	wind	0.0021	74	wadi	0.0014
9	frost	0.0067	42	hun	0.0021	75	vote	0.0014
10	snowing	0.0062	43	fricourt	0.0021	76	pudding	0.0013
11	dugout	0.0056	44	mametz	0.0021	77	needle	0.0013
12	wood	0.0053	45	ribemont	0.0020	78	aunty	0.0013
13	arish	0.0050	46	patient	0.0020	79	duckboard	0.0013
14	ice	0.0048	47	delville	0.0020	80	longueval	0.0013
15	foot	0.0045	48	thick	0.0020	81	armentiere	0.0012
16	camel	0.0038	49	thaw	0.0020	82	misty	0.0012
17	parcel	0.0037	50	le	0.0020	83	evacuated	0.0012
18	bitterly	0.0033	51	amien	0.0019	84	fatigue	0.0012
19	fler	0.0033	52	sleet	0.0019	85	cleaned	0.0011
20	ground	0.0032	53	bath	0.0018	86	football	0.0011
21	freezing	0.0031	54	harness	0.0018	87	houpline	0.0011
22	hard	0.0030	55	sunny	0.0018	88	track	0.0011
23	rum	0.0030	56	bazentin	0.0017	89	sar	0.0011
24	dump	0.0029	57	stayed	0.0017	90	stove	0.0011
25	stable	0.0028	58	tonight	0.0017	91	buire	0.0011
26	muddy	0.0027	59	bernafay	0.0017	92	boche	0.0011
27	frosty	0.0026	60	slush	0.0017	93	falling	0.0011
28	blanket	0.0025	61	eve	0.0017	94	dry	0.0011
29	desert	0.0025	62	gaza	0.0016	95	duck	0.0011
30	foggy	0.0025	63	heavily	0.0016	96	slippery	0.0011
31	trench	0.0024	64	issue	0.0016	97	viv	0.0011
32	rafa	0.0024	65	inch	0.0016	98	havre	0.0011
33	albert	0.0024	66	mazar	0.0016	99	coldest	0.0011

Table F.8: Top 99 terms for the *White Christmas* topic with their probabilities.

After the Armistice

rank	term	beta	rank	term	beta	rank	term	beta
1	train	0.0088	34	aussie	0.0025	67	park	0.0016
2	ship	0.0070	35	picture	0.0025	68	club	0.0016
3	fine	0.0065	36	australia	0.0025	69	trip	0.0016
4	town	0.0062	37	city	0.0023	70	road	0.0016
5	boat	0.0060	38	girl	0.0023	71	band	0.0016
6	sea	0.0052	39	office	0.0023	72	run	0.0016
7	london	0.0049	40	class	0.0023	73	harbour	0.0016
8	hotel	0.0048	41	billet	0.0022	74	mountain	0.0016
9	home	0.0048	42	le	0.0022	75	breakfast	0.0016
10	evening	0.0046	43	germany	0.0022	76	hall	0.0016
11	de	0.0045	44	noon	0.0022	77	supper	0.0016
12	dinner	0.0045	45	passed	0.0022	78	lady	0.0015
13	deck	0.0044	46	ashore	0.0022	79	spent	0.0015
14	port	0.0039	47	hut	0.0022	80	franc	0.0015
15	met	0.0037	48	concert	0.0022	81	meet	0.0015
16	pm	0.0036	49	aboard	0.0021	82	played	0.0014
17	afternoon	0.0036	50	armistice	0.0020	83	digger	0.0014
18	paris	0.0035	51	book	0.0020	84	melbourne	0.0014
19	walk	0.0035	52	charleroi	0.0020	85	journey	0.0014
20	car	0.0035	53	theatre	0.0020	86	left	0.0014
21	troop	0.0035	54	cold	0.0020	87	tram	0.0014
22	lunch	0.0034	55	engine	0.0018	88	visit	0.0014
23	leave	0.0033	56	lorry	0.0018	89	coal	0.0014
24	person	0.0033	57	read	0.0018	90	american	0.0014
25	house	0.0032	58	la	0.0018	91	rue	0.0014
26	bed	0.0032	59	france	0.0018	92	general	0.0013
27	walked	0.0031	60	snow	0.0018	93	german	0.0013
28	board	0.0031	61	crowd	0.0018	94	civilian	0.0013
29	dance	0.0030	62	christmas	0.0017	95	wrote	0.0013
30	street	0.0028	63	seat	0.0017	96	colombo	0.0013
31	tea	0.0027	64	peace	0.0017	97	weather	0.0013
32	war	0.0027	65	card	0.0017	98	lovely	0.0013
33	visited	0.0025	66	pass	0.0016	99	england	0.0013

Table F.9: Top 99 terms for the *After the Armistice* topic with their probabilities.

Home Again

rank	term	beta	rank	term	beta	rank	term	beta
1	home	0.0306	34	piano	0.0029	67	deliver	0.0015
2	meet	0.0158	35	card	0.0027	68	nice	0.0015
3	boat	0.0128	36	middle	0.0026	69	minute	0.0015
4	pm	0.0110	37	rain	0.0025	70	car	0.0015
5	play	0.0095	38	arrive	0.0023	71	hot	0.0014
6	tea	0.0091	39	meeting	0.0023	72	beach	0.0014
7	ring	0.0088	40	boundary	0.0023	73	auntie	0.0014
8	catch	0.0078	41	hour	0.0022	74	lopping	0.0014
9	bed	0.0076	42	read	0.0022	75	motor	0.0014
10	manly	0.0062	43	chat	0.0022	76	afternoon	0.0014
11	walk	0.0057	44	night	0.0022	77	leave	0.0013
12	music	0.0055	45	time	0.0022	78	sleep	0.0013
13	town	0.0055	46	spend	0.0021	79	quay	0.0013
14	mum	0.0052	47	girl	0.0020	80	chri	0.0013
15	tram	0.0051	48	morning	0.0020	81	photo	0.0013
16	dad	0.0049	49	dinner	0.0020	82	call	0.0013
17	george	0.0047	50	shop	0.0019	83	natal	0.0013
18	garden	0.0046	51	dave	0.0019	84	suit	0.0013
19	paddock	0.0044	52	job	0.0019	85	odd	0.0013
20	sit	0.0042	53	lopped	0.0019	86	told	0.0012
21	elli	0.0042	54	flat	0.0018	87	finish	0.0012
22	miss	0.0040	55	train	0.0018	88	billiard	0.0012
23	tickle	0.0039	56	band	0.0017	89	win	0.0012
24	wrote	0.0038	57	picture	0.0017	90	late	0.0012
25	dick	0.0037	58	cut	0.0017	91	happy	0.0012
26	dine	0.0037	59	wharf	0.0017	92	wheeler	0.0012
27	talk	0.0032	60	visit	0.0017	93	letter	0.0011
28	roy	0.0032	61	top	0.0016	94	dance	0.0011
29	swim	0.0031	62	wait	0.0016	95	dee	0.0011
30	drive	0.0031	63	shopping	0.0016	96	rose	0.0011
31	stay	0.0029	64	coat	0.0016	97	wife	0.0011
32	otto	0.0029	65	jack	0.0015	98	darby	0.0011
33	write	0.0029	66	coomoo	0.0015	99	league	0.0011

Table F.10: Top 99 terms for the *Home Again* topic with their probabilities.